Linear Dimensionality Reduction

Practical Machine Learning (CS294-10)

Lecture 6

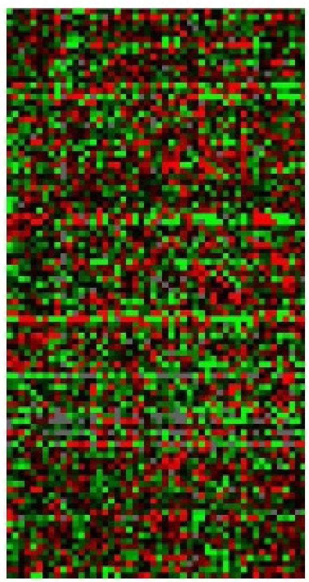
October 16, 2006

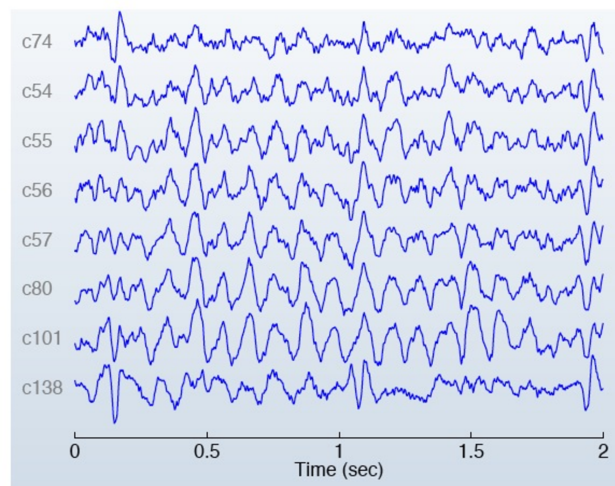
Percy Liang

Lots of high-dimensional noisy data...

Zambian President Levy Mwanawasa has won a second term in office in an election his challenger Michael Sata accused him of rigging, official results showed on Monday.

According to media reports, a pair of hackers said on Saturday that the Firefox Web browser, commonly perceived as the safer and more customizable alternative to market leader Internet Explorer, is critically flawed. A presentation on the flaw was shown during the ToorCon hacker conference in San Diego.



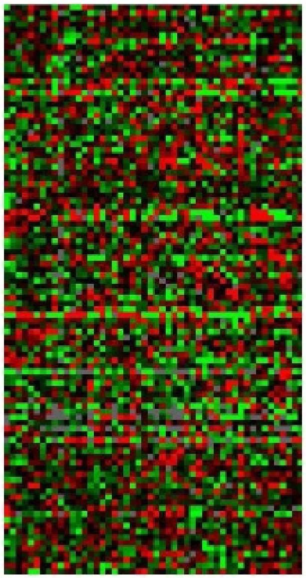


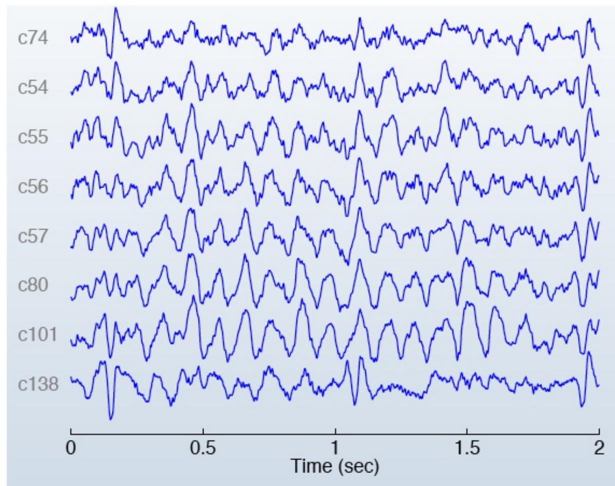
Lots of high-dimensional noisy data...

face images

Zambian President Levy Mwanawasa has won a second term in office in an election his challenger Michael Sata accused him of rigging, official results showed on Monday.

According to media reports, a pair of hackers said on Saturday that the Firefox Web browser, commonly perceived as the safer and more customizable alternative to market leader Internet Explorer, is critically flawed. A presentation on the flaw was shown during the ToorCon hacker conference in San Diego.





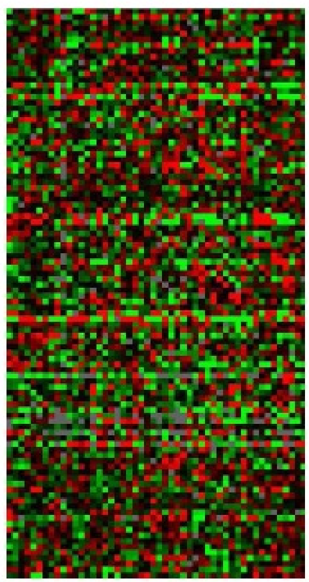
Lots of high-dimensional noisy data...

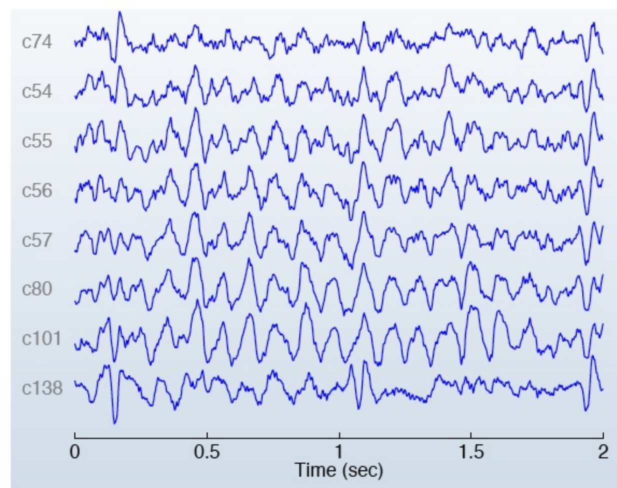
face images

Zambian President Levy Mwanawasa has won a second term in office in an election his challenger Michael Sata accused him of rigging, official results showed on Monday.

According to media reports, a pair of hackers said on Saturday that the Firefox Web browser, commonly perceived as the safer and more customizable alternative to market leader Internet Explorer, is critically flawed. A presentation on the flaw was shown during the ToorCon hacker conference in San Diego.

documents





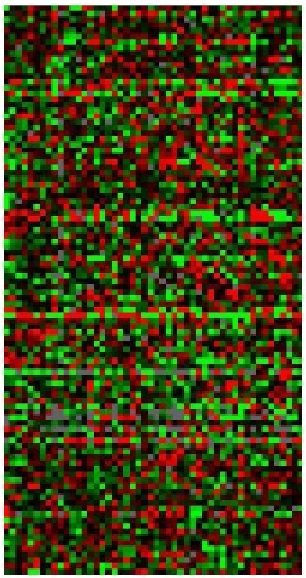
Lots of high-dimensional noisy data...

face images

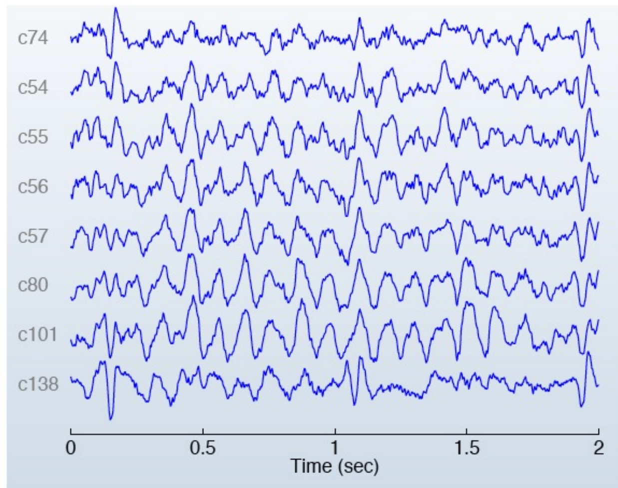
Zambian President Levy Mwanawasa has won a second term in office in an election his challenger Michael Sata accused him of rigging, official results showed on Monday.

According to media reports, a pair of hackers said on Saturday that the Firefox Web browser, commonly perceived as the safer and more customizable alternative to market leader Internet Explorer, is critically flawed. A presentation on the flaw was shown during the ToorCon hacker conference in San Diego.

documents



gene expression data



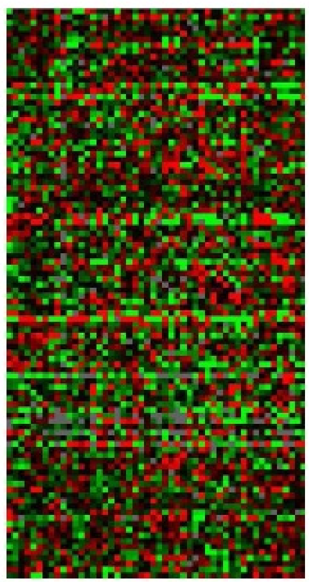
Lots of high-dimensional noisy data...

face images

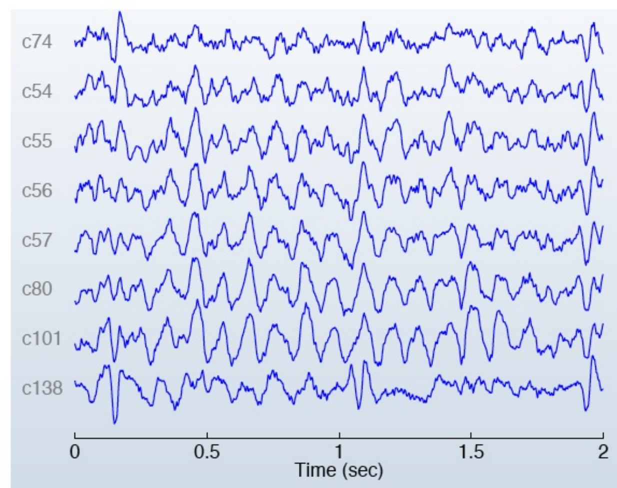
Zambian President Levy Mwanawasa has won a second term in office in an election his challenger Michael Sata accused him of rigging, official results showed on Monday.

According to media reports, a pair of hackers said on Saturday that the Firefox Web browser, commonly perceived as the safer and more customizable alternative to market leader Internet Explorer, is critically flawed. A presentation on the flaw was shown during the ToorCon hacker conference in San Diego.

documents



gene expression data



MEG readings

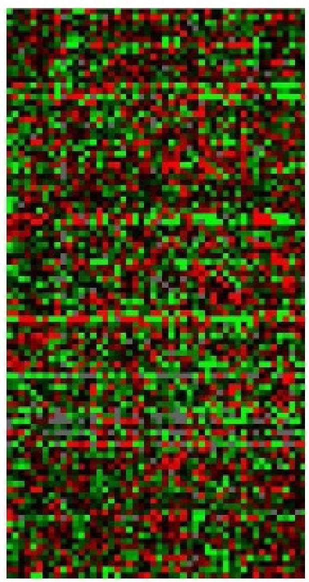
Lots of high-dimensional noisy data...

face images

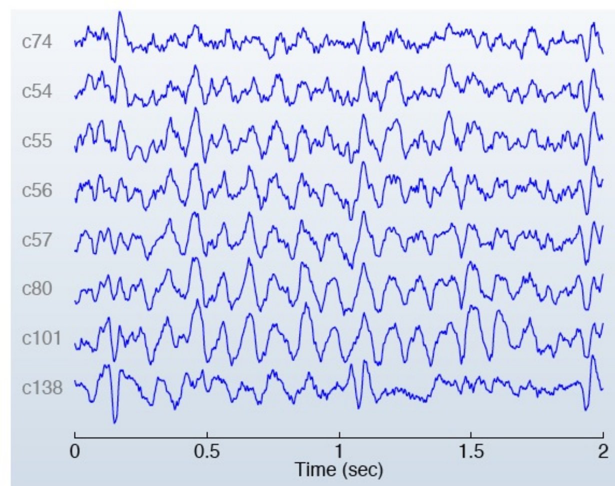
Zambian President Levy Mwanawasa has won a second term in office in an election his challenger Michael Sata accused him of rigging, official results showed on Monday.

According to media reports, a pair of hackers said on Saturday that the Firefox Web browser, commonly perceived as the safer and more customizable alternative to market leader Internet Explorer, is critically flawed. A presentation on the flaw was shown during the ToorCon hacker conference in San Diego.

documents



gene expression data



MEG readings

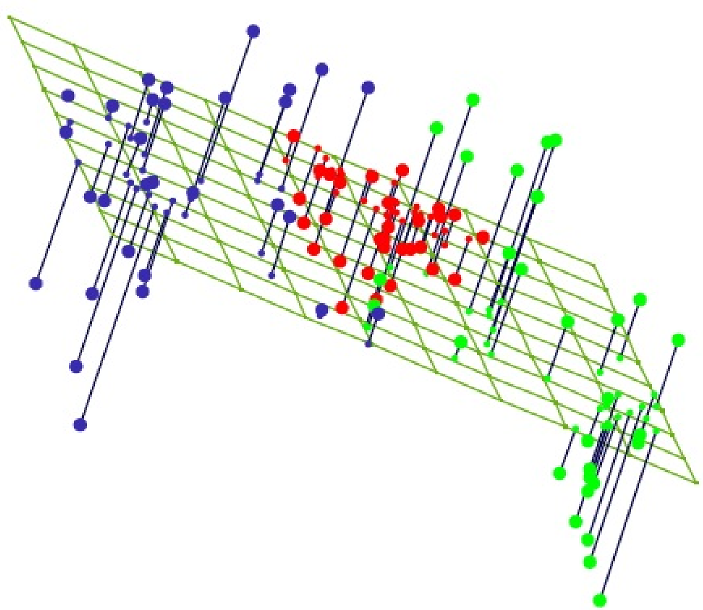
Goal: find a useful representation of data

Basic idea of linear dimensionality reduction

Represent each face as a high-dimensional vector $\mathbf{x} \in \mathbb{R}^{361}$

Basic idea of linear dimensionality reduction

Represent each face as a high-dimensional vector $\mathbf{x} \in \mathbb{R}^{361}$



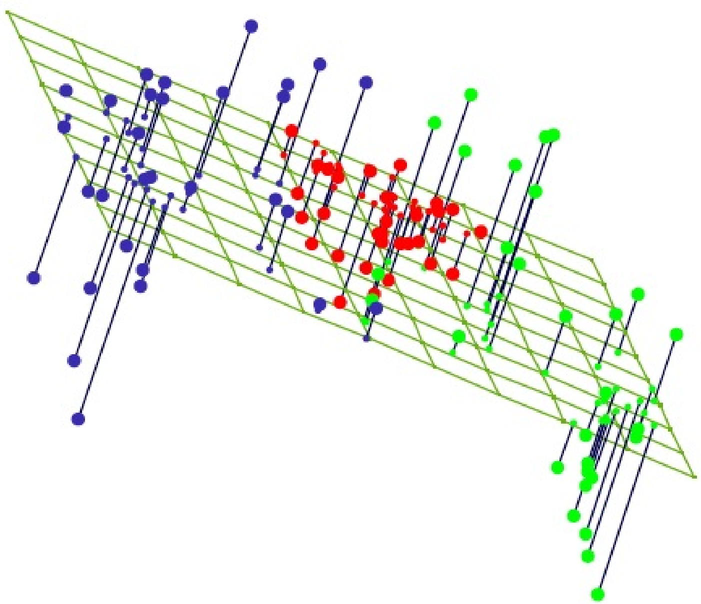
$$\mathbf{x} \in \mathbb{R}^{361}$$

$$\big\downarrow \; \mathbf{z} = \mathbf{U}^T \mathbf{x}$$

$$\mathbf{z} \in \mathbb{R}^{10}$$

Basic idea of linear dimensionality reduction

Represent each face as a high-dimensional vector $\mathbf{x} \in \mathbb{R}^{361}$



$$\mathbf{x} \in \mathbb{R}^{361}$$

$$\Big\downarrow\ \mathbf{z} = \mathbf{U}^T\mathbf{x}$$

$$\mathbf{z} \in \mathbb{R}^{10}$$

This setup is the same for
all methods we will talk about today;
the criteria for choosing $\mathbf{U}$
determines the particular algorithm

Motivation and context

Why do dimensionality reduction? $\mathbf{Z} = \mathbf{U}^T \mathbf{X}$

Motivation and context

Why do dimensionality reduction? $\mathbf{Z} = \mathbf{U}^T\mathbf{X}$

- Scientific: understand structure of data (visualization)

Motivation and context

Why do dimensionality reduction? $\mathbf{Z} = \mathbf{U}^T\mathbf{X}$

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization

Motivation and context

Why do dimensionality reduction? $\mathbf{Z} = \mathbf{U}^T \mathbf{X}$

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
- Computational: compress data for efficiency (both time/space)

Motivation and context

Why do dimensionality reduction? $\mathbf{Z} = \mathbf{U}^T\mathbf{X}$

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
- Computational: compress data for efficiency (both time/space)
- Direct: use as a model for anomaly detection

Motivation and context

Why do dimensionality reduction? $\mathbf{z} = \mathbf{U}^T \mathbf{x}$

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
- Computational: compress data for efficiency (both time/space)
- Direct: use as a model for anomaly detection

In the context of this class...

Motivation and context

Why do dimensionality reduction? $\mathbf{Z} = \mathbf{U}^T \mathbf{X}$

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
- Computational: compress data for efficiency (both time/space)
- Direct: use as a model for anomaly detection

In the context of this class...

- Feature selection (three weeks ago)

Motivation and context

Why do dimensionality reduction? $\mathbf{Z} = \mathbf{U}^T\mathbf{X}$

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
- Computational: compress data for efficiency (both time/space)
- Direct: use as a model for anomaly detection

In the context of this class...

- Feature selection (three weeks ago)
- Clustering (last week)

Motivation and context

Why do dimensionality reduction? $\mathbf{Z} = \mathbf{U}^T\mathbf{X}$

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
- Computational: compress data for efficiency (both time/space)
- Direct: use as a model for anomaly detection

In the context of this class...

- Feature selection (three weeks ago)
- Clustering (last week)
- Nonlinear dimensionality reduction (in 4 weeks)

Motivation and context

Why do dimensionality reduction? $\mathbf{Z} = \mathbf{U}^T\mathbf{X}$

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
- Computational: compress data for efficiency (both time/space)
- Direct: use as a model for anomaly detection

In the context of this class...

- Feature selection (three weeks ago)
- Clustering (last week)
- Nonlinear dimensionality reduction (in 4 weeks)

These are mostly <u>unsupervised</u> methods: use only $\mathbf{X}$
Contrast with supervised methods
(classification, regression), where $(\mathbf{X}, \mathbf{Y})$ are given

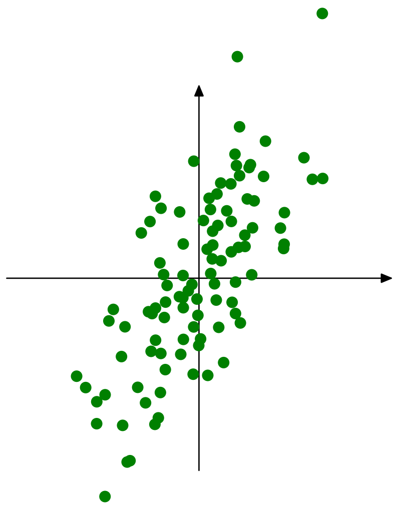
Outline

- Introduction
- Methods
  - Principal component analysis (PCA)
  - Canonical correlation analysis (CCA)
  - Linear discriminant analysis (LDA)
  - Non-negative matrix factorization (NMF)
  - Independent component analysis (ICA)
- Case studies
  - Network anomaly detection
  - Multi-task learning
  - Part-of-speech tagging
  - Brain imaging
- Extensions, related methods, summary

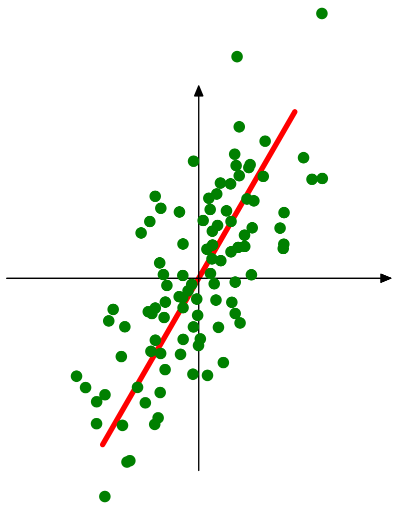
Outline

- Introduction
- Methods
  - Principal component analysis (PCA)
  - Canonical correlation analysis (CCA)
  - Linear discriminant analysis (LDA)
  - Non-negative matrix factorization (NMF)
  - Independent component analysis (ICA)
- Case studies
  - Network anomaly detection
  - Multi-task learning
  - Part-of-speech tagging
  - Brain imaging
- Extensions, related methods, summary

PCA: first principal component

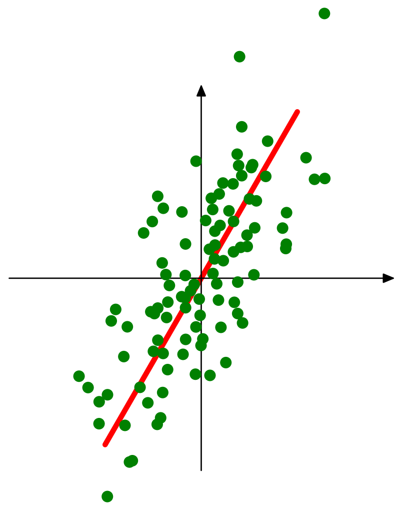


PCA: first principal component



Objective: maximize variance
of projected data

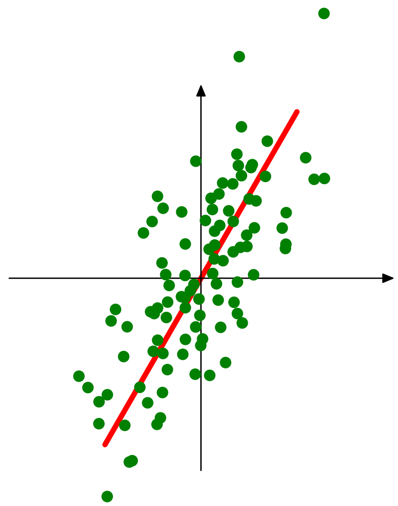
PCA: first principal component



Objective: maximize variance
of projected data

$$= \max_{||\mathbf{u}||=1} \sum_{i=1}^{n} (\underbrace{\mathbf{u}^T \mathbf{x}_i}_{\text{length of projection}})^2$$

PCA: first principal component



Objective: maximize variance of projected data

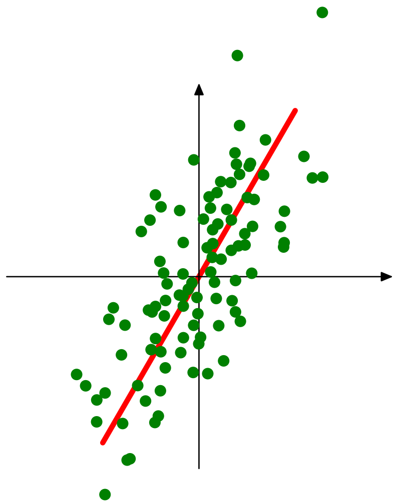
$$= \max_{||\mathbf{u}||=1} \sum_{i=1}^{n} (\underbrace{\mathbf{u}^T \mathbf{x}_i}_{\text{length of projection}})^2$$

$$= \max_{||\mathbf{u}||=1} ||\mathbf{u}^T \mathbf{X}||^2$$

$$\mathbf{X} = \begin{pmatrix} | & & | \\ \mathbf{x}_1 & \dots & \mathbf{x}_n \\ | & & | \end{pmatrix}$$

(assume data is centered at 0)

PCA: first principal component



$$\mathbf{X} = \begin{pmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{pmatrix}$$

(assume data is centered at 0)

Objective: maximize variance of projected data

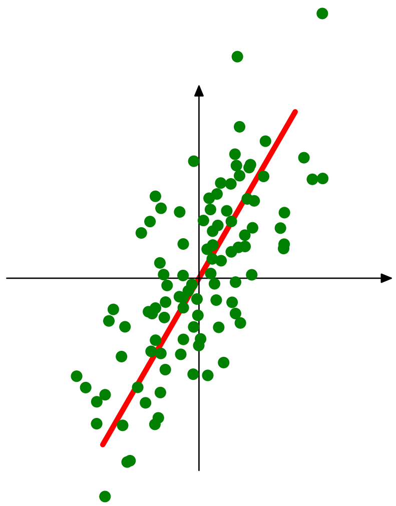
$$= \max_{||\mathbf{u}||=1} \sum_{i=1}^{n} (\underbrace{\mathbf{u}^T \mathbf{x}_i}_{\text{length of projection}})^2$$

$$= \max_{||\mathbf{u}||=1} ||\mathbf{u}^T \mathbf{X}||^2$$

$= \text{largest eigenvalue of } \mathbf{X}\mathbf{X}^T$

(covariance matrix)

PCA: first principal component



$$\mathbf{X} = \begin{pmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{pmatrix}$$

(assume data is centered at 0)

Objective: maximize variance of projected data

$$= \max_{||\mathbf{u}||=1} \sum_{i=1}^{n} (\underbrace{\mathbf{u}^T \mathbf{x}_i}_{\text{length of projection}})^2$$

$$= \max_{||\mathbf{u}||=1} ||\mathbf{u}^T \mathbf{X}||^2$$

$= $ largest eigenvalue of $\mathbf{X}\mathbf{X}^T$

(covariance matrix)

Another perspective:

minimize reconstruction error

$\sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{u}\mathbf{u}^T \mathbf{x}_i||^2$

(similar to least-squares regression?)

All principal components

$$\mathbf{X}_{d\times n} \quad = \quad \mathbf{U}_{d\times d} \qquad \mathbf{Z}_{d\times n}$$

$$\begin{pmatrix} | & & | \\ \mathbf{x}_1 & \ldots & \mathbf{x}_n \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ \mathbf{u}_1 & \ldots & \mathbf{u}_d \\ | & & | \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{z}_1 & \ldots & \mathbf{z}_n \\ | & & | \end{pmatrix}$$

$\mathbf{X}$: data in original representation
$\mathbf{U}$: principal components
$\mathbf{Z}$: data in new representation

All principal components

$$\mathbf{X}_{d\times n} \qquad = \qquad \mathbf{U}_{d\times d} \qquad\qquad \mathbf{Z}_{d\times n}$$

$$\begin{pmatrix} | & & | \\ \mathbf{x}_1 & \dots & \mathbf{x}_n \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ \mathbf{u}_1 & \dots & \mathbf{u}_d \\ | & & | \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{z}_1 & \dots & \mathbf{z}_n \\ | & & | \end{pmatrix}$$

$\mathbf{X}$: data in original representation
$\mathbf{U}$: principal components
$\mathbf{Z}$: data in new representation

- Each $\mathbf{x}_i$ can be expressed by a linear combination of principal components: $\mathbf{x}_i = \sum_{j=1}^{d} z_i^j \mathbf{u}_j$
- Components of projected data are uncorrelated

r principal components

$$\mathbf{X}_{d\times n} \quad \approxeq \quad \mathbf{U}_{d\times r} \quad \mathbf{Z}_{r\times n}$$

$$\begin{pmatrix} | & & | \\ \mathbf{x}_1 & \ldots & \mathbf{x}_n \\ | & & | \end{pmatrix} \approxeq \begin{pmatrix} | & & | \\ \mathbf{u}_1 & \ldots & \mathbf{u}_r \\ | & & | \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{z}_1 & \ldots & \mathbf{z}_n \\ | & & | \end{pmatrix}$$

$\mathbf{X}$: data in original representation
$\mathbf{U}$: principal components
$\mathbf{Z}$: data in new representation

Dimensionality reduction:
keep only the largest r of d eigenvectors

$$\mathbf{x}_i \approxeq \sum_{j=1}^{r} z_i^j \mathbf{u}_j$$

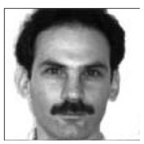
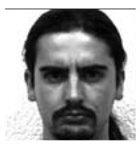
Eigen-faces [Turk, 1991]

Each $\mathbf{x}_i$ is a face image, which is a vector in $\mathbb{R}^d$

d is the number of pixels

Each component $\mathbf{x}_i^j$ is the intensity of the j-th pixel

$$\mathbf{X}_{d\times n} \quad \approxeq \quad \mathbf{U}_{d\times r} \quad \mathbf{Z}_{r\times n}$$

( $\ldots$ ) $\approxeq$ () $\begin{pmatrix} | & & | \\ \mathbf{z}_1 & \ldots & \mathbf{z}_n \\ | & & | \end{pmatrix}$

These faces are from Yale face dataset.

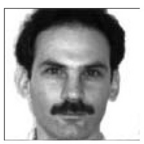
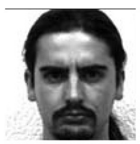
Eigen-faces [Turk, 1991]

Each $\mathbf{x}_i$ is a face image, which is a vector in $\mathbb{R}^d$

d is the number of pixels

Each component $\mathbf{x}_i^j$ is the intensity of the j-th pixel

$$\mathbf{X}_{d\times n} \approxeq \mathbf{U}_{d\times r} \quad \mathbf{Z}_{r\times n}$$

( $\cdots$ ) $\approxeq$ ()

$$\begin{pmatrix} | & & | \\ \mathbf{z}_1 & \cdots & \mathbf{z}_n \\ | & & | \end{pmatrix}$$

Used in image classification.

Individual entries in $\mathbf{z}_i$'s are more meaningful than those in $\mathbf{x}_i$'s.

These faces are from Yale face dataset.

Latent Semantic Analysis [Deerwater, 1990]

Each $\mathbf{x}_i$ is a bag of words, which is a vector in $\mathbb{R}^d$

d is the number of words in the vocabulary

Each component $\mathbf{x}_i^j$ is
 the number of times word j appears in document i

$$\mathbf{X}_{d\times n} \approxeq \mathbf{U}_{d\times r} \; \mathbf{Z}_{r\times n}$$

$$\begin{array}{r} \text{stocks:} \\ \text{chairman:} \\ \text{the:} \\ \cdots \\ \text{wins:} \\ \text{game:} \end{array}\begin{pmatrix} 2 & \cdots & 0 \\ 4 & \cdots & 1 \\ 8 & \cdots & 7 \\ \vdots & \cdots & \vdots \\ 0 & \cdots & 2 \\ 1 & \cdots & 3 \end{pmatrix} \approxeq \begin{pmatrix} 0.4 & \cdots & -0.001 \\ 0.8 & \cdots & 0.03 \\ 0.01 & \cdots & 0.04 \\ \vdots & \cdots & \vdots \\ 0.002 & \cdots & 2.3 \\ 0.003 & \cdots & 1.9 \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{z}_1 & \ldots & \mathbf{z}_n \\ | & & | \end{pmatrix}$$

Latent Semantic Analysis [Deerwater, 1990]

Each $\mathbf{x}_i$ is a bag of words, which is a vector in $\mathbb{R}^d$

d is the number of words in the vocabulary

Each component $\mathbf{x}_i^j$ is

the number of times word j appears in document i

$$\mathbf{X}_{d\times n} \approx \mathbf{U}_{d\times r} \quad \mathbf{Z}_{r\times n}$$

$$\begin{array}{r} \text{stocks:} \\ \text{chairman:} \\ \text{the:} \\ \cdots \\ \text{wins:} \\ \text{game:} \end{array} \begin{pmatrix} 2 & \cdots & 0 \\ 4 & \cdots & 1 \\ 8 & \cdots & 7 \\ \vdots & \cdots & \vdots \\ 0 & \cdots & 2 \\ 1 & \cdots & 3 \end{pmatrix} \approx \begin{pmatrix} 0.4 & \cdots & \text{-}0.001 \\ 0.8 & \cdots & 0.03 \\ 0.01 & \cdots & 0.04 \\ \vdots & \cdots & \vdots \\ 0.002 & \cdots & 2.3 \\ 0.003 & \cdots & 1.9 \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{z}_1 & \ldots & \mathbf{z}_n \\ | & & | \end{pmatrix}$$

Useful in information retrieval.

Eigen-documents gets at notion of semantics.

How to measure similarity between two documents?

$\mathbf{x}_1, \mathbf{x}_2$ versus $\mathbf{z}_1, \mathbf{z}_2$

Computing PCA

- Two ways of generating principal components:
  - Eigendecomposition: $\mathbf{X}\mathbf{X}^T = \mathbf{U}\Lambda\mathbf{U}^T$
  - Singular value decomposition: $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$
- Algorithm:
  - Center data so that $\sum_{i=1}^{n} \mathbf{x}_i = 0$
  - Run SVD (which is one line in R):

```
decomp <- svd(X, r)
```

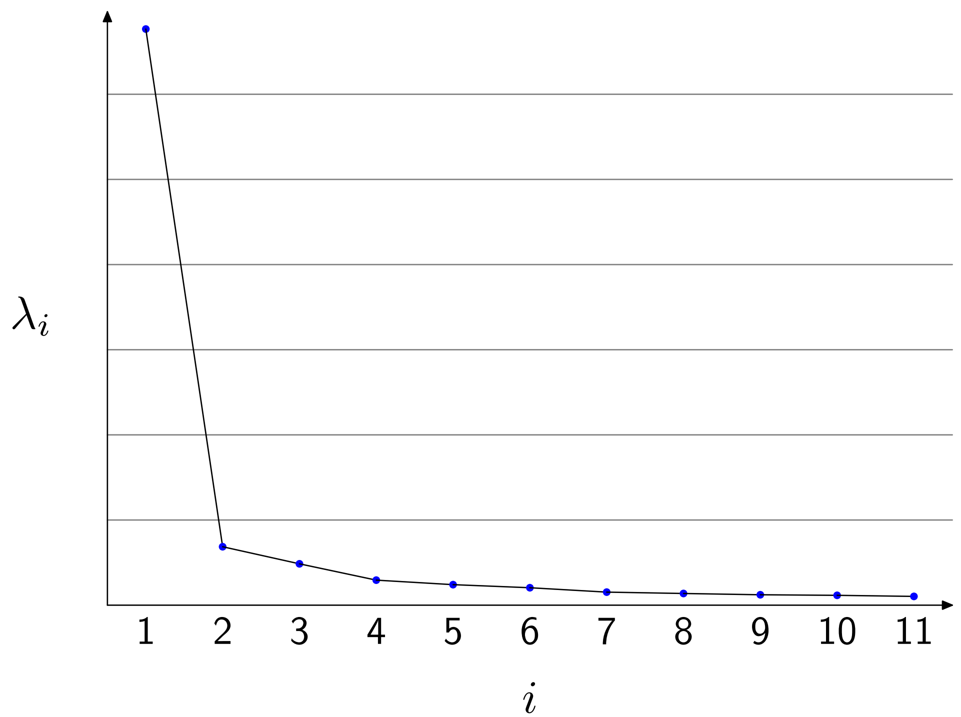
`decomp$u` are principal components
`decomp$d**2` are eigenvalues

How many principal components?

- Similar to question of "How many clusters?"
- Magnitude of eigenvalues indicate percentage of variance captured.

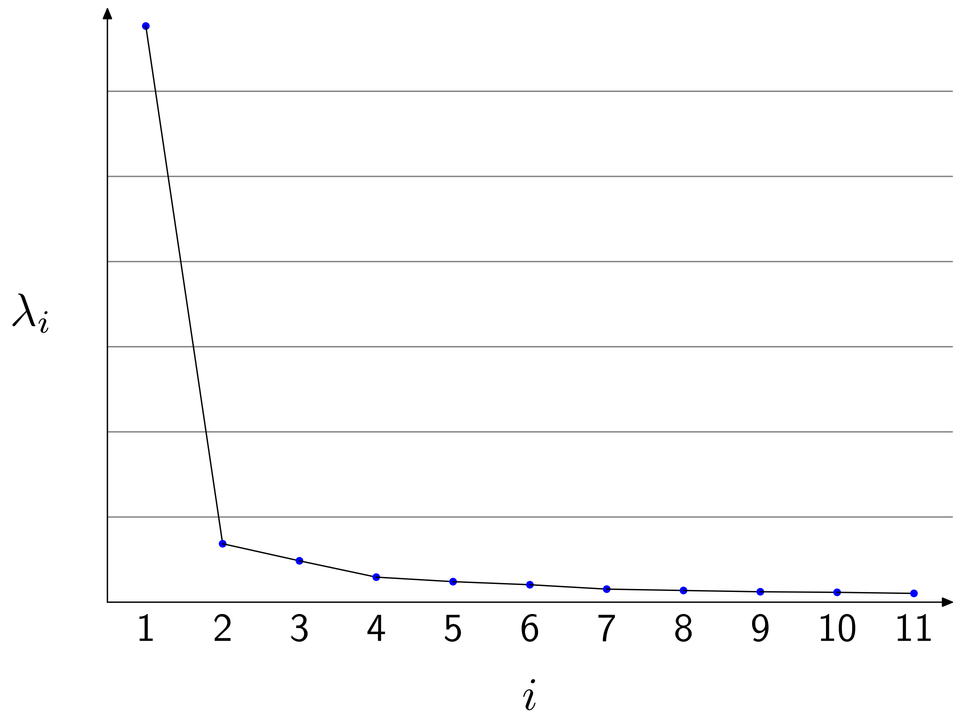
How many principal components?

- Similar to question of "How many clusters?"
- Magnitude of eigenvalues indicate percentage of variance captured.
- Eigenvalues on a face image dataset:



How many principal components?

- Similar to question of "How many clusters?"
- Magnitude of eigenvalues indicate percentage of variance captured.
- Eigenvalues on a face image dataset:



- Eigenvalues drop off sharply, so don't need that many.
- But variance isn't everything...

What if the data doesn't live in a subspace?

- Ideal case: data lies in low-dimensional subspace plus Gaussian noise

What if the data doesn't live in a subspace?

- Ideal case: data lies in low-dimensional subspace plus Gaussian noise
- A hypothetical example:
  - Original data is 100-dimensional
  - True manifold of data is 5-dimensional but lives in a 8-dimensional subspace
  - PCA can just find the 8-dimensional subspace, which still reduces redundancy

What if the data doesn't live in a subspace?

- Ideal case: data lies in low-dimensional subspace plus Gaussian noise
- A hypothetical example:
  - Original data is 100-dimensional
  - True manifold of data is 5-dimensional but lives in a 8-dimensional subspace
  - PCA can just find the 8-dimensional subspace, which still reduces redundancy
- A cool technique: random projections
  - Randomly project data onto $O(\log n)$ dimensions
  - Pairwise distances preserved with high probability
  - Much more efficient than PCA

PCA summary

- Intuition: Capture variance of data
  Minimize reconstruction error
- Algorithm: eigenvalue problem
- Simple to use
- Applications: eigen-faces, eigen-documents,
  eigen-genes, etc.

Outline

- Introduction
- Methods
  - Principal component analysis (PCA)
  - Canonical correlation analysis (CCA)
  - Linear discriminant analysis (LDA)
  - Non-negative matrix factorization (NMF)
  - Independent component analysis (ICA)
- Case studies
  - Network anomaly detection
  - Multi-task learning
  - Part-of-speech tagging
  - Brain imaging
- Extensions, related methods, summary

Motivation for CCA [Hotelling, 1936]

Often, each data point actually consists of many views...

- Image retrieval: for each image, have the following:
  - Pixels (or other visual features)
  - Text around the image

Motivation for CCA [Hotelling, 1936]

Often, each data point actually consists of many views. . .

- Image retrieval: for each image, have the following:
  - Pixels (or other visual features)
  - Text around the image
- Genomics: for each gene, have the following:
  - Gene expression in DNA microarray
  - Position on genome
  - Chemical reactions catalyzed in metabolic pathways

Motivation for CCA [Hotelling, 1936]

Often, each data point actually consists of many views...

- Image retrieval: for each image, have the following:
  - Pixels (or other visual features)
  - Text around the image
- Genomics: for each gene, have the following:
  - Gene expression in DNA microarray
  - Position on genome
  - Chemical reactions catalyzed in metabolic pathways

Goal: reduce the dimensionality of the views <u>jointly</u>

From variance to correlation

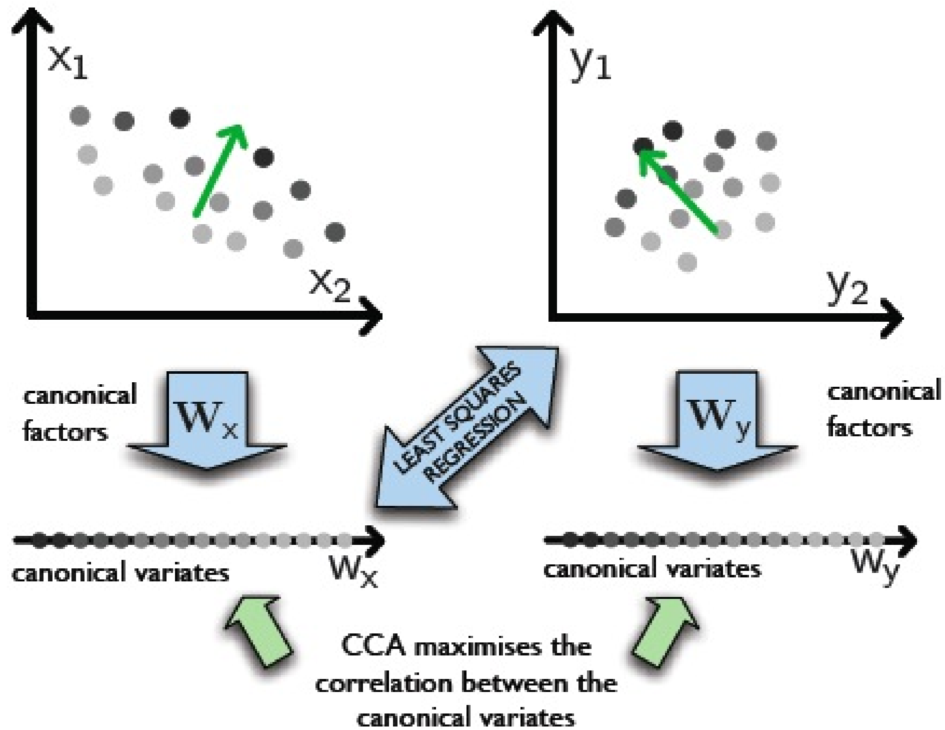
PCA: find $\mathbf{u}$ to maximize variance $\hat{\mathbb{E}}(\mathbf{u}^T\mathbf{x})^2$

CCA: find $(\mathbf{u}, \mathbf{v})$ to maximize correlation $\widehat{\text{corr}}(\mathbf{u}^T\mathbf{x})(\mathbf{v}^T\mathbf{y})$

From variance to correlation

PCA: find $\mathbf{u}$ to maximize variance $\hat{\mathbb{E}}(\mathbf{u}^T\mathbf{x})^2$

CCA: find $(\mathbf{u}, \mathbf{v})$ to maximize correlation $\widehat{\text{corr}}(\mathbf{u}^T\mathbf{x})(\mathbf{v}^T\mathbf{y})$

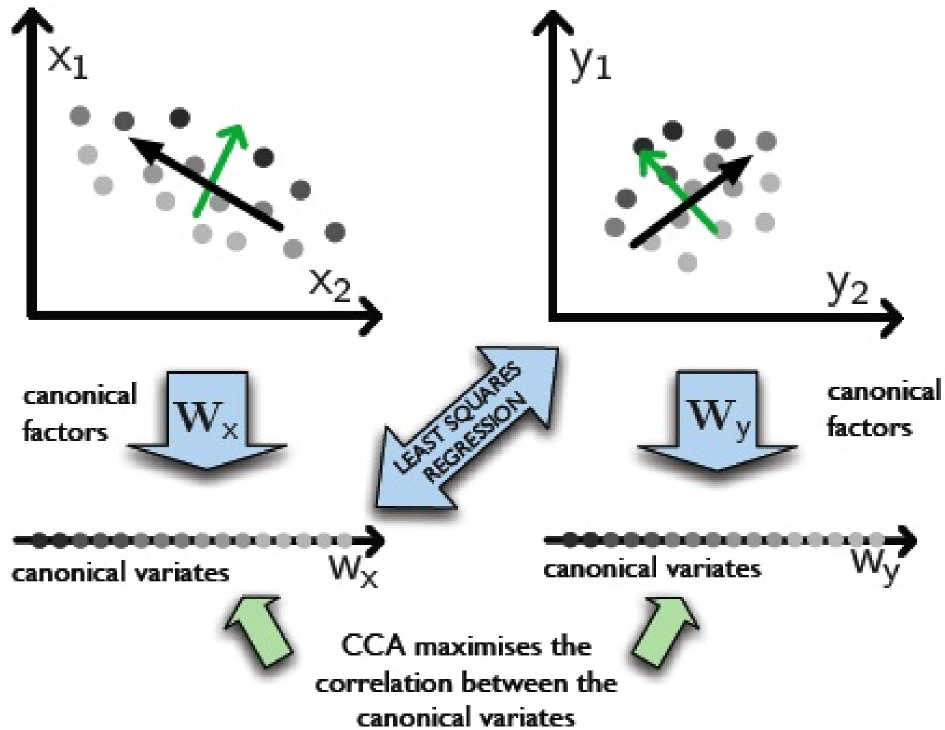


CCA directions (green)

From variance to correlation

PCA: find $\mathbf{u}$ to maximize variance $\hat{\mathbb{E}}(\mathbf{u}^T\mathbf{x})^2$

CCA: find $(\mathbf{u}, \mathbf{v})$ to maximize correlation $\widehat{\text{corr}}(\mathbf{u}^T\mathbf{x})(\mathbf{v}^T\mathbf{y})$

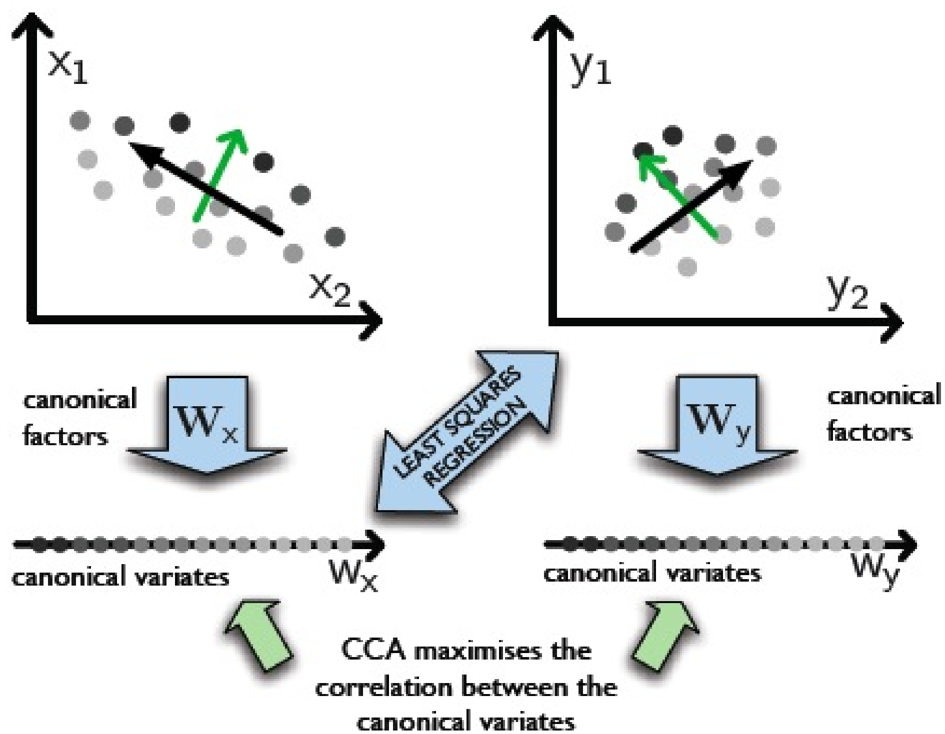


CCA directions (green) PCA directions (black)

From variance to correlation

PCA: find $\mathbf{u}$ to maximize variance $\hat{\mathbb{E}}(\mathbf{u}^T\mathbf{x})^2$

CCA: find $(\mathbf{u}, \mathbf{v})$ to maximize correlation $\widehat{\text{corr}}(\mathbf{u}^T\mathbf{x})(\mathbf{v}^T\mathbf{y})$



CCA directions (green) PCA directions (black)

Doing PCA separately on each view does not take advantage of relationship between two views.

CCA objective function

Objective: maximize correlation between projected views

CCA objective function

Objective: maximize correlation between projected views

$$= \max_{\mathbf{u},\mathbf{v}} \widehat{\mathsf{corr}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y}) = \max_{\mathbf{u},\mathbf{v}} \frac{\widehat{\mathsf{cov}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y})}{\sqrt{\widehat{\mathsf{var}}(\mathbf{u}^T\mathbf{x})}\sqrt{\widehat{\mathsf{var}}(\mathbf{v}^T\mathbf{y})}}$$

CCA objective function

Objective: maximize correlation between projected views

$$= \max_{\mathbf{u},\mathbf{v}} \widehat{\mathsf{corr}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y}) = \max_{\mathbf{u},\mathbf{v}} \frac{\widehat{\mathsf{cov}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y})}{\sqrt{\widehat{\mathsf{var}}(\mathbf{u}^T\mathbf{x})}\sqrt{\widehat{\mathsf{var}}(\mathbf{v}^T\mathbf{y})}}$$

$$= \max_{\widehat{\mathsf{var}}(\mathbf{u}^T\mathbf{x})=\widehat{\mathsf{var}}(\mathbf{v}^T\mathbf{y})=1} \widehat{\mathsf{cov}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y})$$

CCA objective function

Objective: maximize correlation between projected views

$$= \max_{\mathbf{u},\mathbf{v}} \widehat{\mathsf{corr}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y}) = \max_{\mathbf{u},\mathbf{v}} \frac{\widehat{\mathsf{cov}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y})}{\sqrt{\widehat{\mathsf{var}}(\mathbf{u}^T\mathbf{x})}\sqrt{\widehat{\mathsf{var}}(\mathbf{v}^T\mathbf{y})}}$$

$$= \max_{\widehat{\mathsf{var}}(\mathbf{u}^T\mathbf{x})=\widehat{\mathsf{var}}(\mathbf{v}^T\mathbf{y})=1} \widehat{\mathsf{cov}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y})$$

$$= \max_{||\mathbf{u}^T\mathbf{X}||=||\mathbf{v}^T\mathbf{Y}||=1} \sum_{i=1}^{n} (\mathbf{u}^T\mathbf{x}_i)(\mathbf{v}^T\mathbf{y}_i)$$

CCA objective function

Objective: maximize correlation between projected views

$$= \max_{\mathbf{u},\mathbf{v}} \widehat{\mathsf{corr}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y}) = \max_{\mathbf{u},\mathbf{v}} \frac{\widehat{\mathsf{cov}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y})}{\sqrt{\widehat{\mathsf{var}}(\mathbf{u}^T\mathbf{x})}\sqrt{\widehat{\mathsf{var}}(\mathbf{v}^T\mathbf{y})}}$$

$$= \max_{\widehat{\mathsf{var}}(\mathbf{u}^T\mathbf{x})=\widehat{\mathsf{var}}(\mathbf{v}^T\mathbf{y})=1} \widehat{\mathsf{cov}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y})$$

$$= \max_{||\mathbf{u}^T\mathbf{X}||=||\mathbf{v}^T\mathbf{Y}||=1} \sum_{i=1}^{n} (\mathbf{u}^T\mathbf{x}_i)(\mathbf{v}^T\mathbf{y}_i)$$

$$= \max_{||\mathbf{u}^T\mathbf{X}||=||\mathbf{v}^T\mathbf{Y}||=1} \mathbf{u}^T\mathbf{X}\mathbf{Y}^T\mathbf{v}$$

CCA objective function

Objective: maximize correlation between projected views

$$= \max_{\mathbf{u},\mathbf{v}} \widehat{\mathsf{corr}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y}) = \max_{\mathbf{u},\mathbf{v}} \frac{\widehat{\mathsf{cov}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y})}{\sqrt{\widehat{\mathsf{var}}(\mathbf{u}^T\mathbf{x})}\sqrt{\widehat{\mathsf{var}}(\mathbf{v}^T\mathbf{y})}}$$

$$= \max_{\widehat{\mathsf{var}}(\mathbf{u}^T\mathbf{x})=\widehat{\mathsf{var}}(\mathbf{v}^T\mathbf{y})=1} \widehat{\mathsf{cov}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y})$$

$$= \max_{||\mathbf{u}^T\mathbf{X}||=||\mathbf{v}^T\mathbf{Y}||=1} \sum_{i=1}^{n} (\mathbf{u}^T\mathbf{x}_i)(\mathbf{v}^T\mathbf{y}_i)$$

$$= \max_{||\mathbf{u}^T\mathbf{X}||=||\mathbf{v}^T\mathbf{Y}||=1} \mathbf{u}^T\mathbf{X}\mathbf{Y}^T\mathbf{v}$$

$=$ largest generalized eigenvalue λ given by

$$\begin{pmatrix} 0 & \mathbf{X}\mathbf{Y}^T \\ \mathbf{Y}\mathbf{X}^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{X}\mathbf{X}^T & 0 \\ 0 & \mathbf{Y}\mathbf{Y}^T \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix},$$

which reduces to an ordinary eigenvalue problem.

CCA objective function

Objective: maximize correlation between projected views

$$= \max_{\mathbf{u},\mathbf{v}} \widehat{\mathsf{corr}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y}) = \max_{\mathbf{u},\mathbf{v}} \frac{\widehat{\mathsf{cov}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y})}{\sqrt{\widehat{\mathsf{var}}(\mathbf{u}^T\mathbf{x})}\sqrt{\widehat{\mathsf{var}}(\mathbf{v}^T\mathbf{y})}}$$

$$= \max_{\widehat{\mathsf{var}}(\mathbf{u}^T\mathbf{x})=\widehat{\mathsf{var}}(\mathbf{v}^T\mathbf{y})=1} \widehat{\mathsf{cov}}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y})$$

$$= \max_{||\mathbf{u}^T\mathbf{X}||=||\mathbf{v}^T\mathbf{Y}||=1} \sum_{i=1}^{n} (\mathbf{u}^T\mathbf{x}_i)(\mathbf{v}^T\mathbf{y}_i)$$

$$= \max_{||\mathbf{u}^T\mathbf{X}||=||\mathbf{v}^T\mathbf{Y}||=1} \mathbf{u}^T\mathbf{X}\mathbf{Y}^T\mathbf{v}$$

$=$ largest generalized eigenvalue λ given by

$$\begin{pmatrix} 0 & \mathbf{X}\mathbf{Y}^T \\ \mathbf{Y}\mathbf{X}^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{X}\mathbf{X}^T & 0 \\ 0 & \mathbf{Y}\mathbf{Y}^T \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix},$$

which reduces to an ordinary eigenvalue problem.

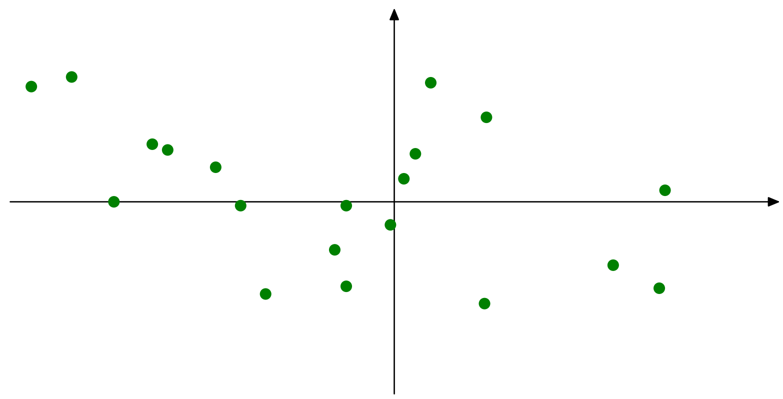
Note: canonical components $\mathbf{u}, \mathbf{v}$ are invariant to affine transformation of $\mathbf{X}, \mathbf{Y}$

Outline

- Introduction
- Methods
  - Principal component analysis (PCA)
  - Canonical correlation analysis (CCA)
  - Linear discriminant analysis (LDA)
  - Non-negative matrix factorization (NMF)
  - Independent component analysis (ICA)
- Case studies
  - Network anomaly detection
  - Multi-task learning
  - Part-of-speech tagging
  - Brain imaging
- Extensions, related methods, summary

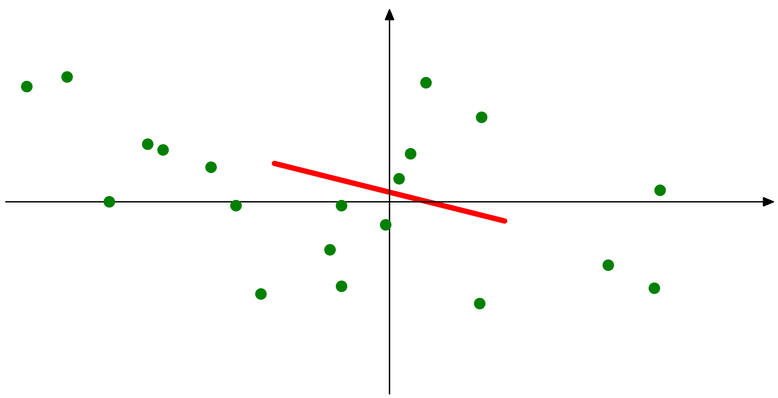
Motivation for LDA [Fisher, 1936]

What is the best linear projection?



Motivation for LDA [Fisher, 1936]

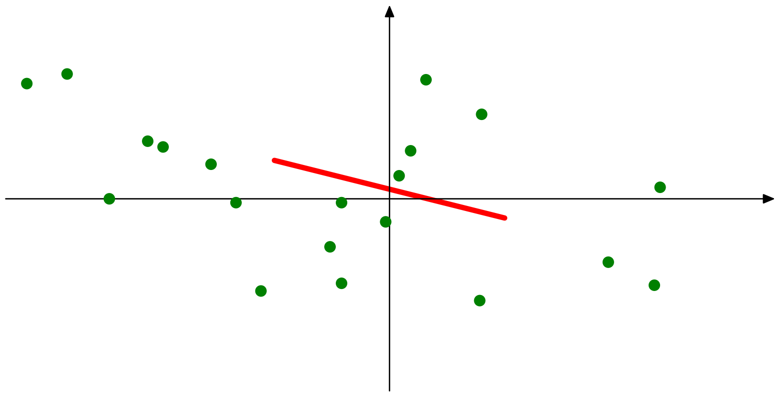
What is the best linear projection?



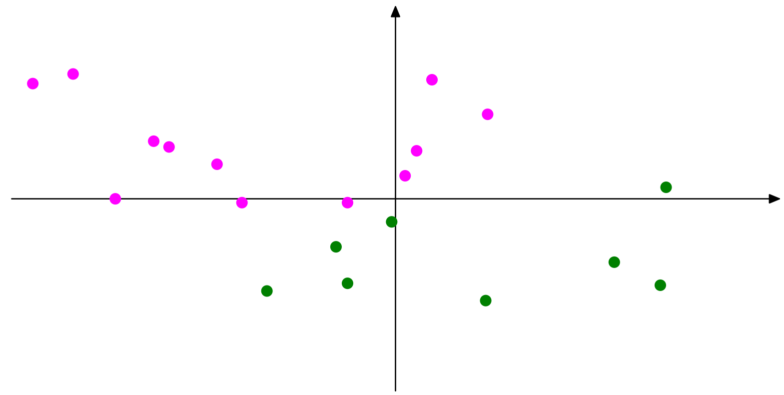
PCA solution

Motivation for LDA [Fisher, 1936]

What is the best linear projection with these labels?

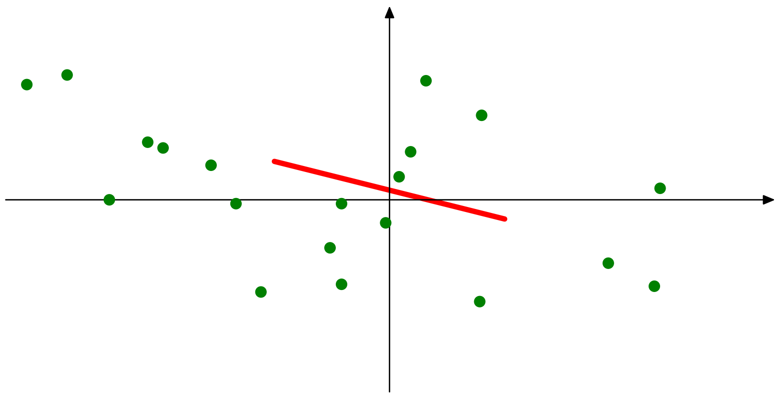


PCA solution

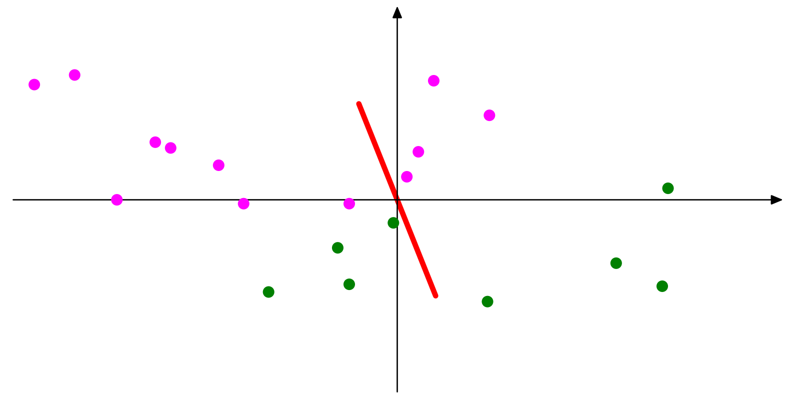


Motivation for LDA [Fisher, 1936]

What is the best linear projection with these labels?



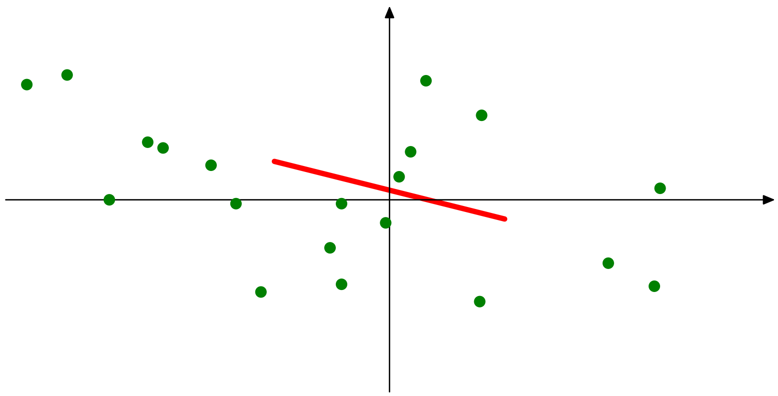
PCA solution



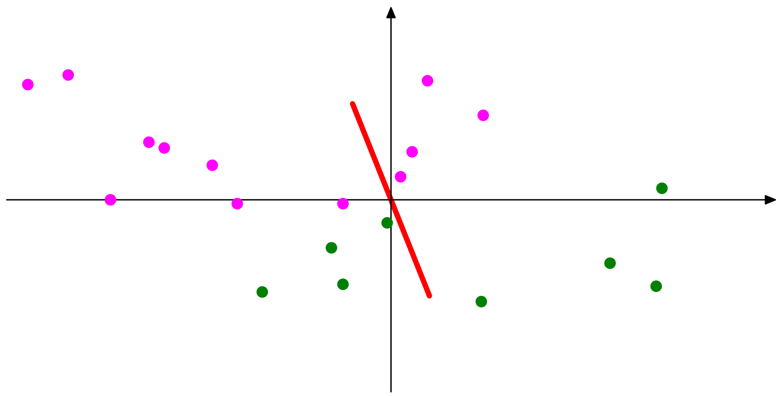
LDA solution

Motivation for LDA [Fisher, 1936]

What is the best linear projection with these labels?



PCA solution



LDA solution

Goal: reduce the dimensionality <u>given labels</u>

Idea: want projection to maximize overall interclass variance
relative to intraclass variance

LDA objective function

Global mean: $\mu = \sum_i \mathbf{x}_i$ $\mathbf{X}_g = (\mathbf{x}_1 - \mu, \ldots, \mathbf{x}_n - \mu)$

Class mean: $\mu_y = \sum_{i:\mathbf{y}_i = y} \mathbf{x}_i$ $\mathbf{X}_c = (\mathbf{x}_1 - \mu_{\mathbf{y}_1}, \ldots, \mathbf{x}_n - \mu_{\mathbf{y}_n})$

LDA objective function

Global mean: $\mu = \sum_i \mathbf{x}_i$ $\quad$ $\mathbf{X}_g = (\mathbf{x}_1 - \mu, \ldots, \mathbf{x}_n - \mu)$

Class mean: $\mu_y = \sum_{i:\mathbf{y}_i = y} \mathbf{x}_i$ $\quad$ $\mathbf{X}_c = (\mathbf{x}_1 - \mu_{\mathbf{y}_1}, \ldots, \mathbf{x}_n - \mu_{\mathbf{y}_n})$

Objective: maximize $\frac{\text{total variance}}{\text{intraclass variance}} = \frac{\text{interclass variance}}{\text{intraclass variance}} + 1$

LDA objective function

Global mean: $\mu = \sum_i \mathbf{x}_i$ $\quad \mathbf{X}_g = (\mathbf{x}_1 - \mu, \ldots, \mathbf{x}_n - \mu)$

Class mean: $\mu_y = \sum_{i:\mathbf{y}_i = y} \mathbf{x}_i$ $\quad \mathbf{X}_c = (\mathbf{x}_1 - \mu_{\mathbf{y}_1}, \ldots, \mathbf{x}_n - \mu_{\mathbf{y}_n})$

Objective: maximize $\frac{\text{total variance}}{\text{intraclass variance}} = \frac{\text{interclass variance}}{\text{intraclass variance}} + 1$

$$= \max_{\mathbf{u}} \frac{\sum_{i=1}^{n} (\mathbf{u}^T(\mathbf{x}_i - \mu))^2}{\sum_{i=1}^{n} (\mathbf{u}^T(\mathbf{x}_i - \mu_{\mathbf{y}_i}))^2}$$

LDA objective function

Global mean: $\mu = \sum_i \mathbf{x}_i$ $\qquad \mathbf{X}_g = (\mathbf{x}_1 - \mu, \ldots, \mathbf{x}_n - \mu)$

Class mean: $\mu_y = \sum_{i:\mathbf{y}_i = y} \mathbf{x}_i$ $\quad \mathbf{X}_c = (\mathbf{x}_1 - \mu_{\mathbf{y}_1}, \ldots, \mathbf{x}_n - \mu_{\mathbf{y}_n})$

Objective: maximize $\frac{\text{total variance}}{\text{intraclass variance}} = \frac{\text{interclass variance}}{\text{intraclass variance}} + 1$

$$= \max_{\mathbf{u}} \frac{\sum_{i=1}^n (\mathbf{u}^T(\mathbf{x}_i - \mu))^2}{\sum_{i=1}^n (\mathbf{u}^T(\mathbf{x}_i - \mu_{\mathbf{y}_i}))^2}$$

$$= \max_{||\mathbf{u}^T \mathbf{X}_c|| = 1} \sum_{i=1}^n (\mathbf{u}^T(\mathbf{x}_i - \mu))^2$$

LDA objective function

Global mean: $\mu = \sum_i \mathbf{x}_i$ $\quad\quad$ $\mathbf{X}_g = (\mathbf{x}_1 - \mu, \ldots, \mathbf{x}_n - \mu)$

Class mean: $\mu_y = \sum_{i:\mathbf{y}_i = y} \mathbf{x}_i$ $\quad$ $\mathbf{X}_c = (\mathbf{x}_1 - \mu_{\mathbf{y}_1}, \ldots, \mathbf{x}_n - \mu_{\mathbf{y}_n})$

Objective: maximize $\frac{\text{total variance}}{\text{intraclass variance}} = \frac{\text{interclass variance}}{\text{intraclass variance}} + 1$

$$= \max_{\mathbf{u}} \frac{\sum_{i=1}^{n} (\mathbf{u}^T(\mathbf{x}_i - \mu))^2}{\sum_{i=1}^{n} (\mathbf{u}^T(\mathbf{x}_i - \mu_{\mathbf{y}_i}))^2}$$

$$= \max_{||\mathbf{u}^T \mathbf{X}_c|| = 1} \sum_{i=1}^{n} (\mathbf{u}^T(\mathbf{x}_i - \mu))^2$$

$$= \max_{||\mathbf{u}^T \mathbf{X}_c|| = 1} \mathbf{u}^T \mathbf{X}_g \mathbf{X}_g^T \mathbf{u}$$

LDA objective function

Global mean: $\mu = \sum_i \mathbf{x}_i$ $\quad\quad$ $\mathbf{X}_g = (\mathbf{x}_1 - \mu, \ldots, \mathbf{x}_n - \mu)$

Class mean: $\mu_y = \sum_{i:\mathbf{y}_i = y} \mathbf{x}_i$ $\quad$ $\mathbf{X}_c = (\mathbf{x}_1 - \mu_{\mathbf{y}_1}, \ldots, \mathbf{x}_n - \mu_{\mathbf{y}_n})$

Objective: maximize $\frac{\text{total variance}}{\text{intraclass variance}} = \frac{\text{interclass variance}}{\text{intraclass variance}} + 1$

$$= \max_{\mathbf{u}} \frac{\sum_{i=1}^n (\mathbf{u}^T(\mathbf{x}_i - \mu))^2}{\sum_{i=1}^n (\mathbf{u}^T(\mathbf{x}_i - \mu_{\mathbf{y}_i}))^2}$$

$$= \max_{||\mathbf{u}^T \mathbf{X}_c|| = 1} \sum_{i=1}^n (\mathbf{u}^T(\mathbf{x}_i - \mu))^2$$

$$= \max_{||\mathbf{u}^T \mathbf{X}_c|| = 1} \mathbf{u}^T \mathbf{X}_g \mathbf{X}_g^T \mathbf{u}$$

$=$ largest generalized eigenvalue λ given by

$$(\mathbf{X}_g \mathbf{X}_g^T)\mathbf{u} = \lambda(\mathbf{X}_c \mathbf{X}_c^T)\mathbf{u}.$$

Summary so far

- Recall $\mathbf{Z} \approxeq \mathbf{U}^T\mathbf{X}$; criteria for $\mathbf{U}$:
  - PCA: maximize variance
  - CCA: maximize correlation
  - LDA: maximize $\frac{\text{interclass variance}}{\text{intraclass variance}}$
- All these methods reduce to solving generalized eigenvalue problems
- Next (NMF, ICA):
  more complex criteria for $\mathbf{U}$

Outline

- Introduction
- Methods
  - Principal component analysis (PCA)
  - Canonical correlation analysis (CCA)
  - Linear discriminant analysis (LDA)
  - Non-negative matrix factorization (NMF)
  - Independent component analysis (ICA)
- Case studies
  - Network anomaly detection
  - Multi-task learning
  - Part-of-speech tagging
  - Brain imaging
- Extensions, related methods, summary

Motivation for NMF [Paatero, '94; Lee, '99]

Back to basic PCA setting (single view, no labels)

$$\mathbf{X}_{d \times n} \quad \approxeq \quad \mathbf{U}_{d \times r} \quad \mathbf{Z}_{r \times n}$$

$$\begin{pmatrix} | & & | \\ \mathbf{x}_1 & \dots & \mathbf{x}_n \\ | & & | \end{pmatrix} \approxeq \begin{pmatrix} | & & | \\ \mathbf{u}_1 & \dots & \mathbf{u}_r \\ | & & | \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{z}_1 & \dots & \mathbf{z}_n \\ | & & | \end{pmatrix}$$

$\mathbf{X}$: data in original representation
$\mathbf{U}$: principal components
$\mathbf{Z}$: data in new representation

Motivation for NMF [Paatero, '94; Lee, '99]

Back to basic PCA setting (single view, no labels)

$$\mathbf{X}_{d\times n} \quad \approxeq \quad \mathbf{U}_{d\times r} \quad \mathbf{Z}_{r\times n}$$

$$\begin{pmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{pmatrix} \approxeq \begin{pmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_r \\ | & & | \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{z}_1 & \cdots & \mathbf{z}_n \\ | & & | \end{pmatrix}$$

- Data is not just any arbitrary real vector:
  - Text modeling: each document is a vector of term frequencies
  - Gene expression: each gene is a vector of expression profiles
  - Collaborative filtering: each user is a vector of movie ratings

Motivation for NMF [Paatero, '94; Lee, '99]

Back to basic PCA setting (single view, no labels)

$$\mathbf{X}_{d\times n} \quad \approxeq \quad \mathbf{U}_{d\times r} \quad \mathbf{Z}_{r\times n}$$

$$\begin{pmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{pmatrix} \approxeq \begin{pmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_r \\ | & & | \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{z}_1 & \cdots & \mathbf{z}_n \\ | & & | \end{pmatrix}$$

- Data is not just any arbitrary real vector:
  - Text modeling: each document is a vector of term frequencies
  - Gene expression: each gene is a vector of expression profiles
  - Collaborative filtering: each user is a vector of movie ratings
- Each basis vector $\mathbf{u}_i$ is an "eigen-document/eigen-gene/eigen-user"
- Would like $\mathbf{U}$ and $\mathbf{Z}$ to have only non-negative entries
  so that we can <u>interpret</u> each point as combination of prototypes

Motivation for NMF [Paatero, '94; Lee, '99]

Back to basic PCA setting (single view, no labels)

$$\mathbf{X}_{d\times n} \approxeq \mathbf{U}_{d\times r} \quad \mathbf{Z}_{r\times n}$$

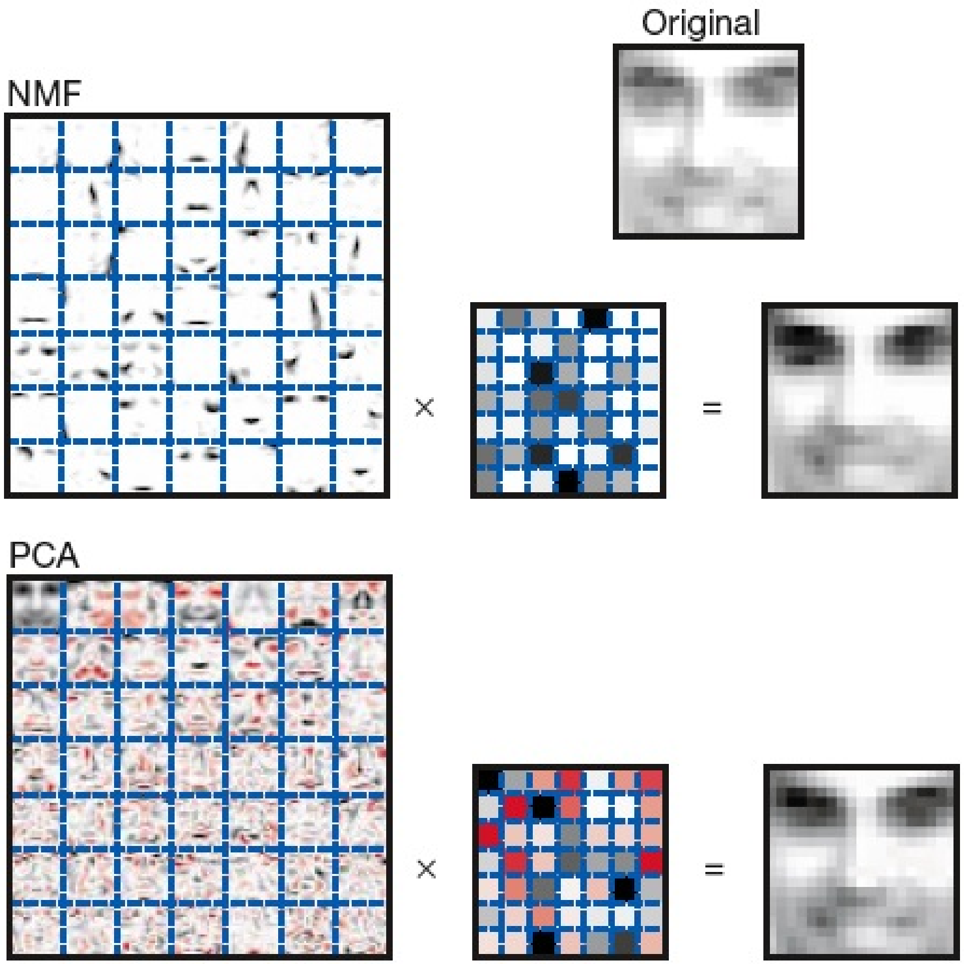
$$\begin{pmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{pmatrix} \approxeq \begin{pmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_r \\ | & & | \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{z}_1 & \cdots & \mathbf{z}_n \\ | & & | \end{pmatrix}$$

- Data is not just any arbitrary real vector:
  - Text modeling: each document is a vector of term frequencies
  - Gene expression: each gene is a vector of expression profiles
  - Collaborative filtering: each user is a vector of movie ratings
- Each basis vector $\mathbf{u}_i$ is an "eigen-document/eigen-gene/eigen-user"
- Would like $\mathbf{U}$ and $\mathbf{Z}$ to have only non-negative entries
  so that we can <u>interpret</u> each point as combination of prototypes

Goal: reduce the dimensionality <u>given non-negativity constraints</u>

Qualitative difference between NMF and PCA

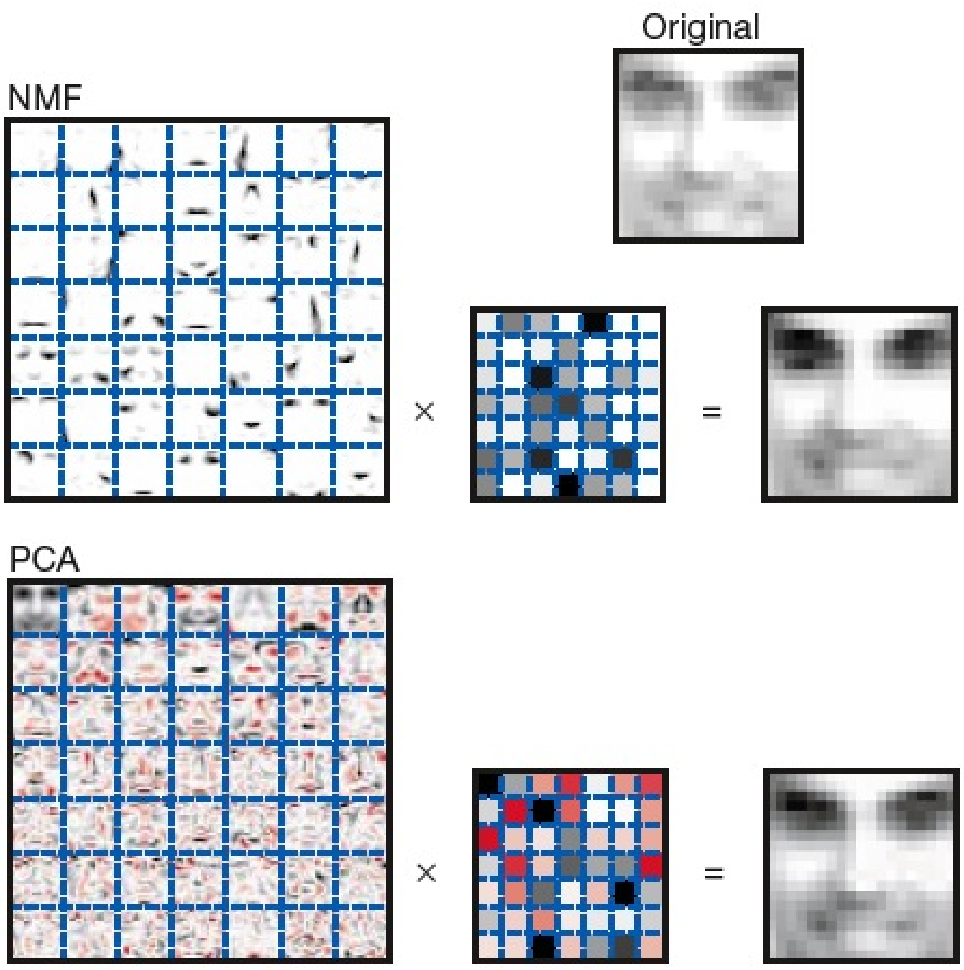
$$\mathbf{x} \approxeq \sum_{j=1}^{r} z_j \mathbf{u}_j$$



- Sum of basis vectors must be (positively) additive ($z_j \geq 0$)
- The basis vectors $\mathbf{u}_i$'s tend to be sparse
- NMF recovers a parts-based representation of $\mathbf{x}$ whereas PCA recovers a holistic representations

Qualitative difference between NMF and PCA

$$\mathbf{x} \approxeq \sum_{j=1}^{r} z_j \mathbf{u}_j$$



- Sum of basis vectors must be (positively) additive ($z_j \geq 0$)
- The basis vectors $\mathbf{u}_i$'s tend to be sparse
- NMF recovers a parts-based representation of $\mathbf{x}$ whereas PCA recovers a holistic representations
- Caveat for images: sparsity depends on proper alignment (remember, representation is still a bag of pixels)

NMF machinery

- Objectives to minimize (all entries in $\mathbf{X}, \mathbf{U}, \mathbf{Z}$ non-negative)
  - Frobenius norm (same as PCA but with non-negativity constraints):
    $||\mathbf{X} - \mathbf{U}\mathbf{Z}||_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{r} (\mathbf{X}_{ji} - (\mathbf{U}\mathbf{Z})_{ji})^2$
  - KL divergence:
    $\mathsf{KL}(\mathbf{X}||\mathbf{U}\mathbf{Z}) = \sum_{i=1}^{n} \sum_{j=1}^{r} \mathbf{X}_{ji} \log \frac{\mathbf{X}_{ji}}{(\mathbf{U}\mathbf{Z})_{ji}} - \mathbf{X}_{ji} + (\mathbf{U}\mathbf{Z})_{ji}$

NMF machinery

- Objectives to minimize (all entries in $\mathbf{X}, \mathbf{U}, \mathbf{Z}$ non-negative)
  - Frobenius norm (same as PCA but with non-negativity constraints):
    $||\mathbf{X} - \mathbf{U}\mathbf{Z}||_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{r} (\mathbf{X}_{ji} - (\mathbf{U}\mathbf{Z})_{ji})^2$
  - KL divergence:
    $\mathsf{KL}(\mathbf{X}||\mathbf{U}\mathbf{Z}) = \sum_{i=1}^{n} \sum_{j=1}^{r} \mathbf{X}_{ji} \log \frac{\mathbf{X}_{ji}}{(\mathbf{U}\mathbf{Z})_{ji}} - \mathbf{X}_{ji} + (\mathbf{U}\mathbf{Z})_{ji}$
- Algorithm
  - Hard non-convex optimization problem:
    could get stuck in local minima, need to worry about initialization
  - Simple/fast multiplicative update rule [Lee & Seung '99, '01]

NMF machinery

- Objectives to minimize (all entries in $\mathbf{X}, \mathbf{U}, \mathbf{Z}$ non-negative)
  - Frobenius norm (same as PCA but with non-negativity constraints):
    $||\mathbf{X} - \mathbf{U}\mathbf{Z}||_F^2 = \sum_{i=1}^n \sum_{j=1}^r (\mathbf{X}_{ji} - (\mathbf{U}\mathbf{Z})_{ji})^2$
  - KL divergence:
    $\mathsf{KL}(\mathbf{X}||\mathbf{U}\mathbf{Z}) = \sum_{i=1}^n \sum_{j=1}^r \mathbf{X}_{ji} \log \frac{\mathbf{X}_{ji}}{(\mathbf{U}\mathbf{Z})_{ji}} - \mathbf{X}_{ji} + (\mathbf{U}\mathbf{Z})_{ji}$
- Algorithm
  - Hard non-convex optimization problem:
    could get stuck in local minima, need to worry about initialization
  - Simple/fast multiplicative update rule [Lee & Seung '99, '01]
- Relationship to other methods
  - Vector quantization: z_j is 1 in exactly one component j
  - Probabilistic latent semantic analysis: equivalent to 2nd objective
  - Latent Dirichlet Allocation: more Bayesian version of pLSI

Outline

- Introduction
- Methods
  - Principal component analysis (PCA)
  - Canonical correlation analysis (CCA)
  - Linear discriminant analysis (LDA)
  - Non-negative matrix factorization (NMF)
  - Independent component analysis (ICA)
- Case studies
  - Network anomaly detection
  - Multi-task learning
  - Part-of-speech tagging
  - Brain imaging
- Extensions, related methods, summary

Motivation for ICA [Herault & Jutten, '86]

$\mathbf{x} = \mathbf{U}\mathbf{z}$

Cocktail party problem:

d people, d microphones, n time steps

Assume: people are speaking independently ($\mathbf{z}$)

acoustics mix linearly through an invertible $\mathbf{U}$

Motivation for ICA [Herault & Jutten, '86]

$\mathbf{x} = \mathbf{U}\mathbf{z}$

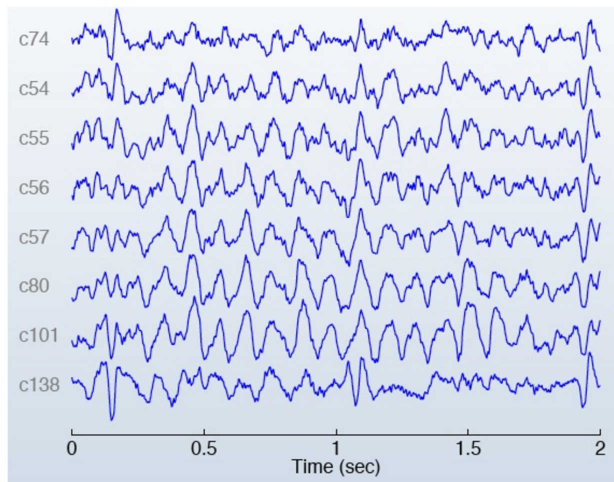
Cocktail party problem:

d people, d microphones, n time steps

Assume: people are speaking independently ($\mathbf{z}$)

acoustics mix linearly through an invertible $\mathbf{U}$

$$\mathbf{X} =$$



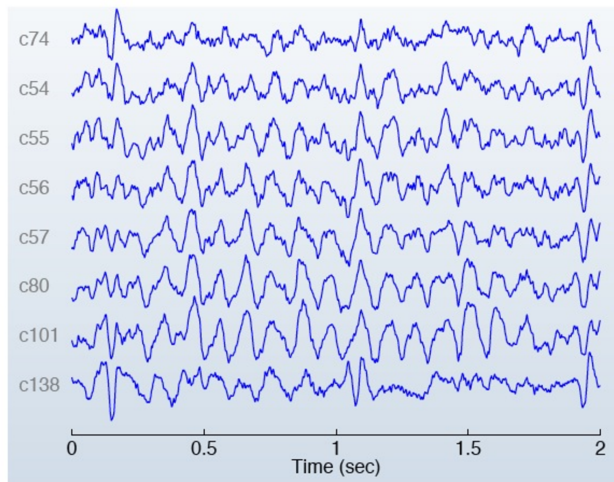
Motivation for ICA [Herault & Jutten, '86]

$\mathbf{x} = \mathbf{U}\mathbf{z}$

Cocktail party problem:

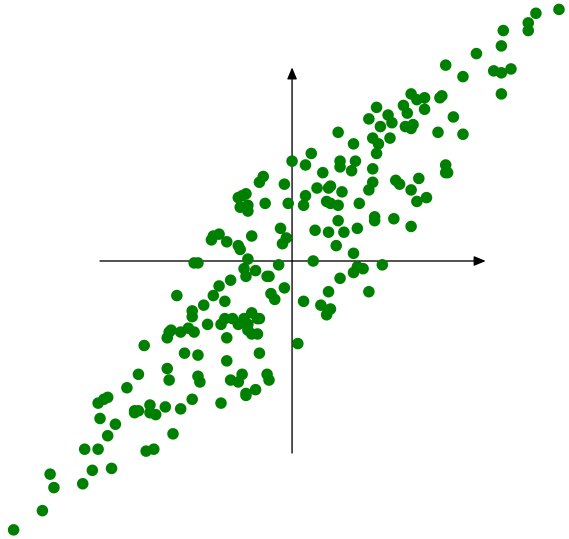
d people, d microphones, n time steps

Assume: people are speaking independently ($\mathbf{z}$)
acoustics mix linearly through an invertible $\mathbf{U}$

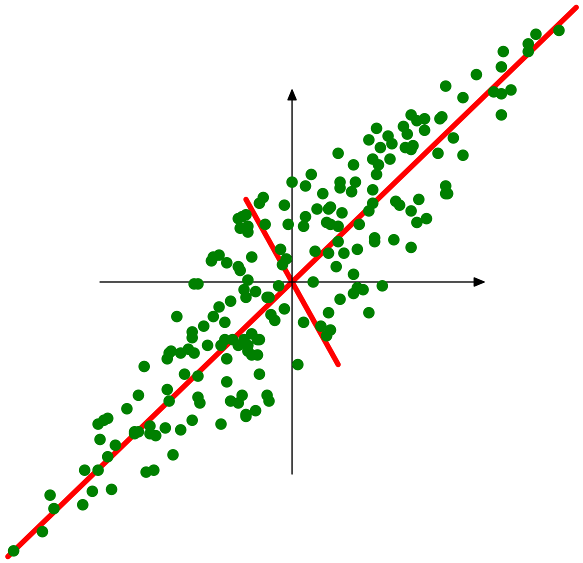
$\mathbf{X} =$ 

Goal: find transformation that makes components of $\mathbf{z}$ as independent as possible

PCA versus ICA

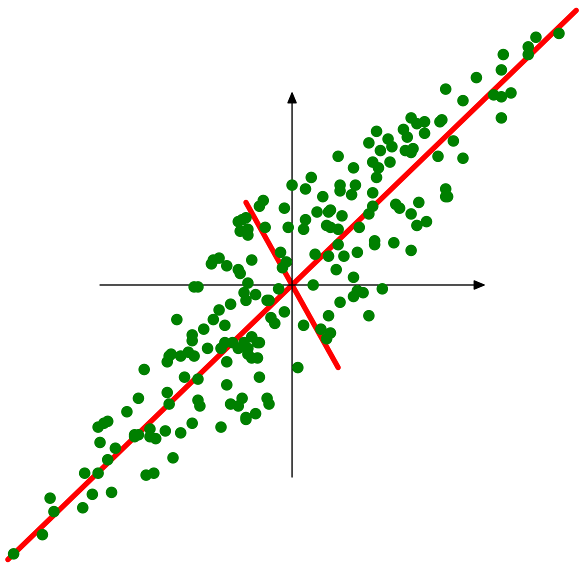


PCA versus ICA

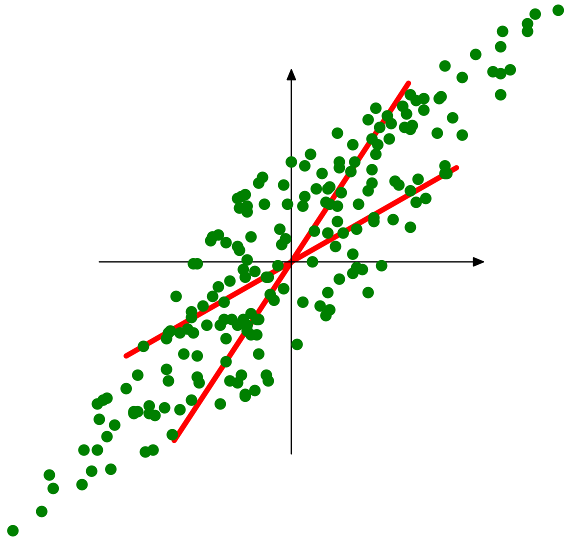


PCA solution

PCA versus ICA

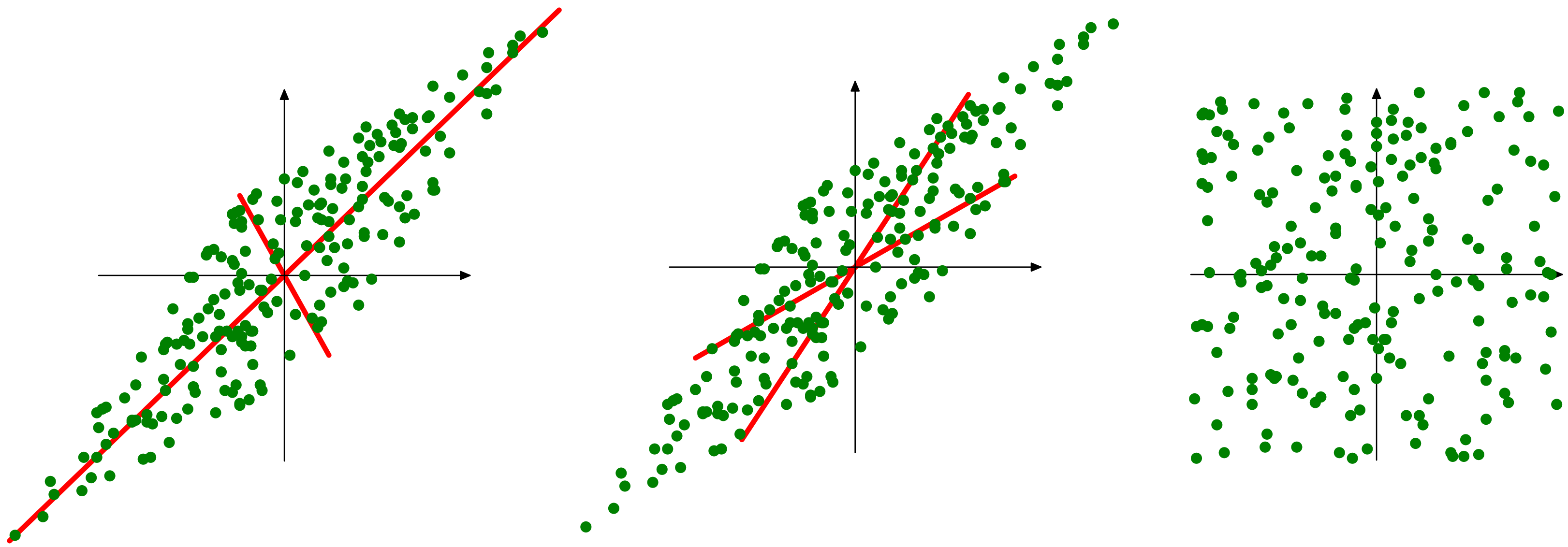


PCA solution



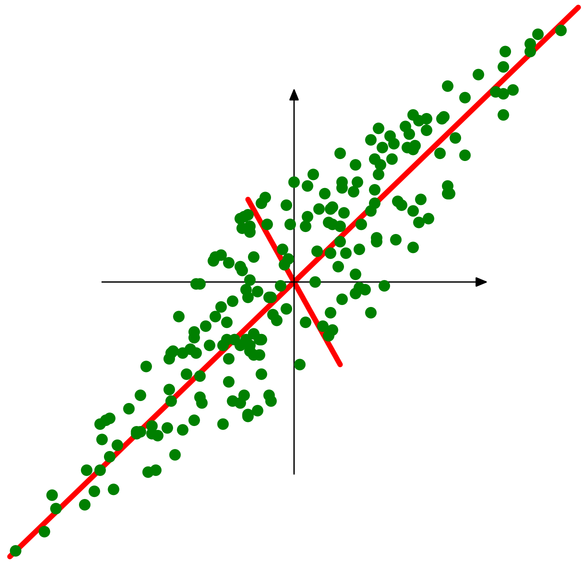
ICA solution

PCA versus ICA

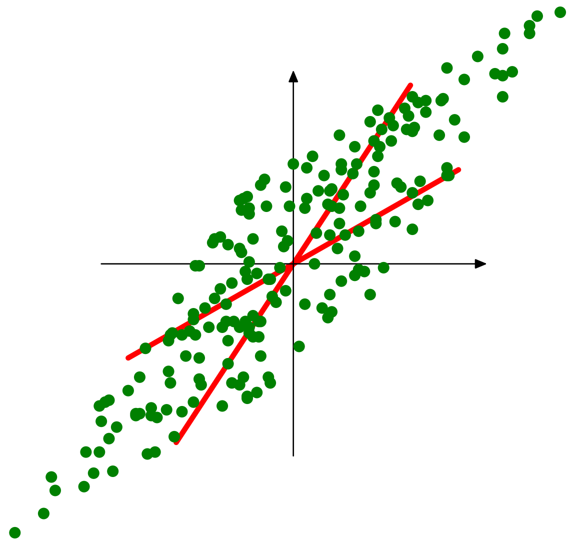


PCA solution ICA solution Original signal

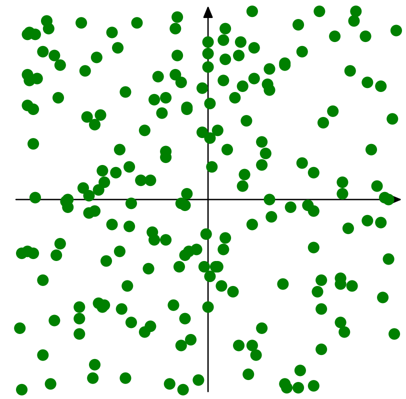
PCA versus ICA



PCA solution

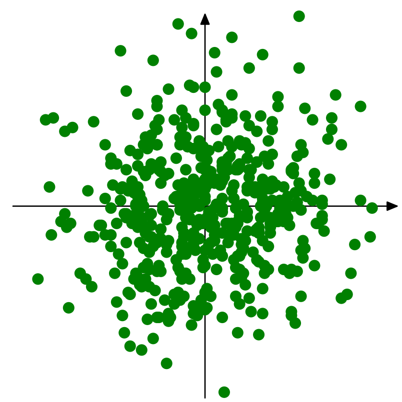


ICA solution

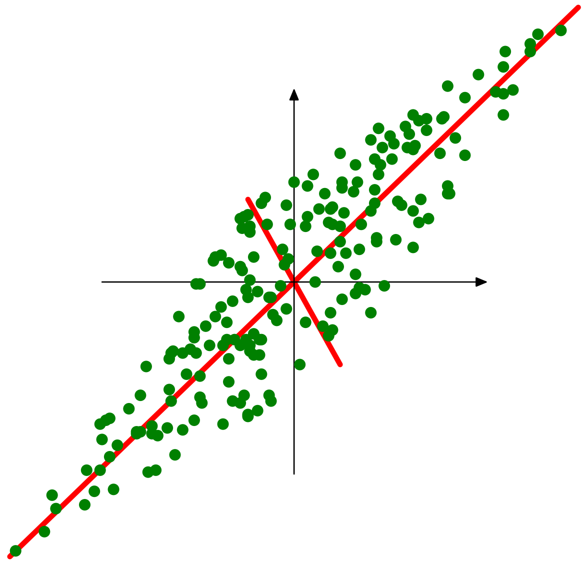


Original signal

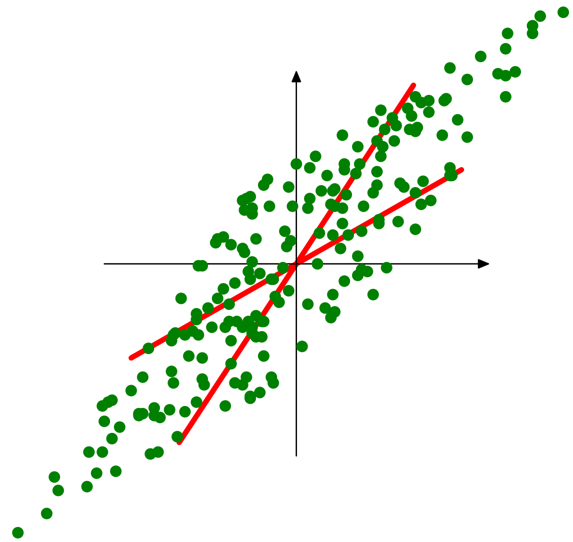
ICA finds independent components; doesn't work if data is Gaussian:



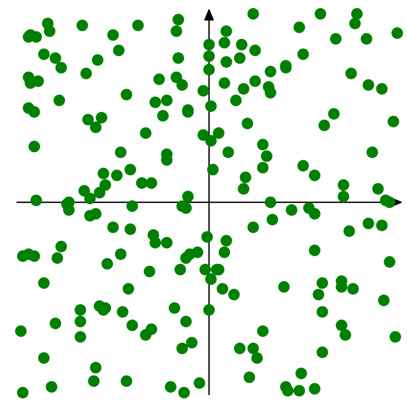
PCA versus ICA



PCA solution

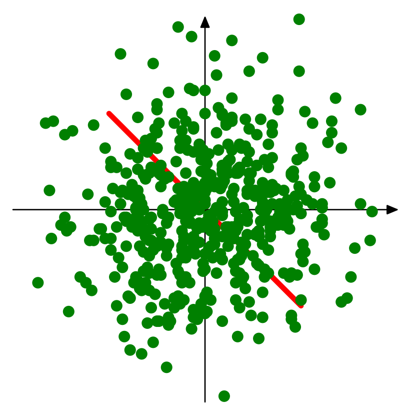


ICA solution



Original signal

ICA finds independent components; doesn't work if data is Gaussian:



?

ICA algorithm

$$\mathbf{x} = \mathbf{U}\mathbf{z}$$

- Preprocessing: whiten data $\mathbf{X}$ with PCA so that components are uncorrelated

ICA algorithm

$$\mathbf{x} = \mathbf{U}\mathbf{z}$$

- Preprocessing: whiten data $\mathbf{X}$ with PCA so that components are uncorrelated
- Find $\mathbf{U}^{-1}$ to maximize independence of $\mathbf{z} = \mathbf{U}^{-1}\mathbf{x}$
- How to measure independence? mutual information, negentropy, non-Gaussianity (e.g., kurtosis)

ICA algorithm

$$\mathbf{x} = \mathbf{U}\mathbf{z}$$

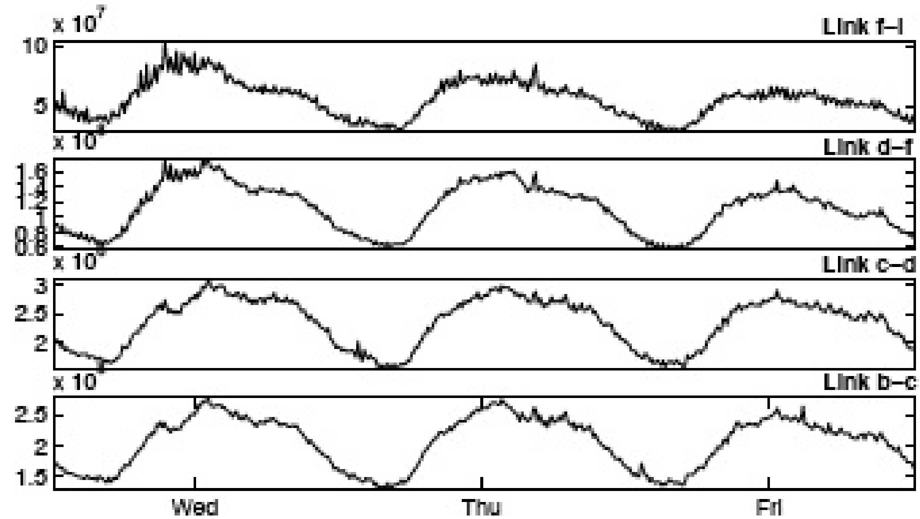
- Preprocessing: whiten data $\mathbf{X}$ with PCA so that components are uncorrelated
- Find $\mathbf{U}^{-1}$ to maximize independence of $\mathbf{z} = \mathbf{U}^{-1}\mathbf{x}$
- How to measure independence? mutual information, negentropy, non-Gaussianity (e.g., kurtosis)
- Hard non-convex optimization
- Methods for solving: fastICA, kernelICA, ProDenICA

Outline

- Introduction
- Methods
  - Principal component analysis (PCA)
  - Canonical correlation analysis (CCA)
  - Linear discriminant analysis (LDA)
  - Non-negative matrix factorization (NMF)
  - Independent component analysis (ICA)
- Case studies
  - Network anomaly detection
  - Multi-task learning
  - Part-of-speech tagging
  - Brain imaging
- Extensions, related methods, summary

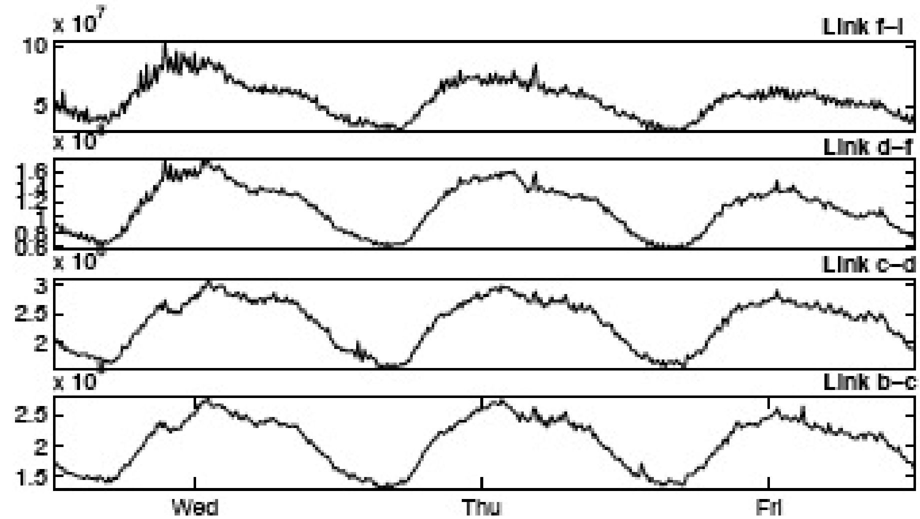
Network anomaly detection [Lakhina, '05]

Raw data: traffic flow on each link in the network during each time interval



Network anomaly detection [Lakhina, '05]

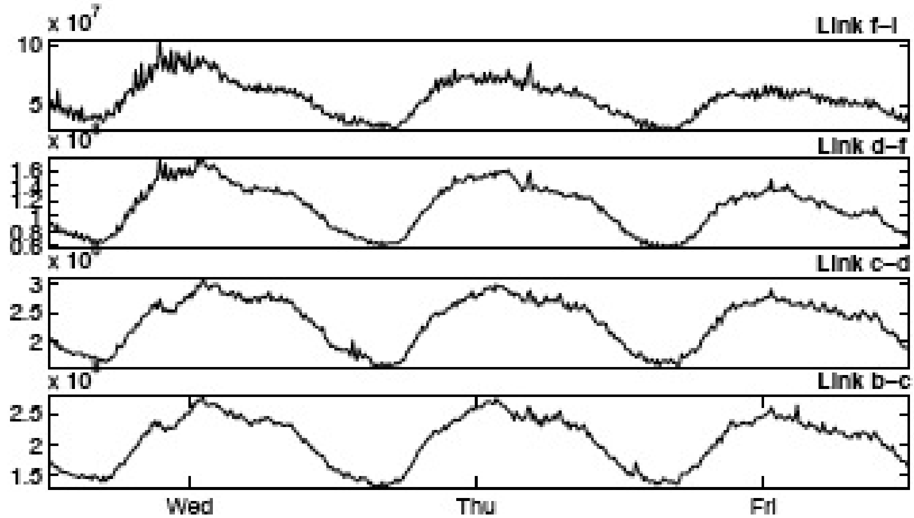
Raw data: traffic flow on each link in the network during each time interval



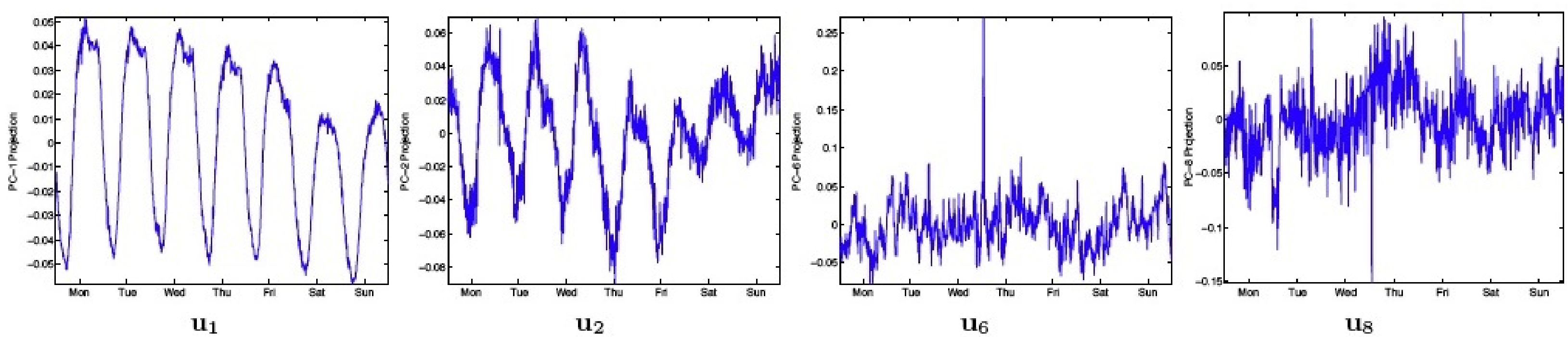
Model assumption: traffic is sum of flows along a few paths
Apply PCA: principal component intuitively represents a path

Network anomaly detection [Lakhina, '05]

Raw data: traffic flow on each link in the network during each time interval



Model assumption: traffic is sum of flows along a few paths
Apply PCA: principal component intuitively represents a path
Anomaly: when traffic deviates from first few principal components



(a) Normal Behavior (b) Anomalous Behavior

Multi-task learning [Ando & Zhang, '05]

Setup:

- Have a set of related tasks (classify documents for various users)
- Each task has a classifier (weights of a linear classifier)
- Want to share structure between classifiers

Multi-task learning [Ando & Zhang, '05]

Setup:

- Have a set of related tasks (classify documents for various users)
- Each task has a classifier (weights of a linear classifier)
- Want to share structure between classifiers

One step of their procedure:
given a set of classifiers $\mathbf{x}_1, \dots, \mathbf{x}_n$,
run PCA to identify shared structure:

$$\mathbf{X} = \begin{pmatrix} | & & | \\ \mathbf{x}_1 & \dots & \mathbf{x}_n \\ | & & | \end{pmatrix} \approxeq \mathbf{U}\mathbf{Z}$$

Each data point is a linear classifier
Each principal component is a eigen-classifier

Unsupervised POS tagging [Schütze, '95]

Part-of-speech (POS) tagging task:

Input:	I	like	reducing	the	dimensionality	of	data	.
Output:	NOUN	VERB	VERB(-ING)	DET	NOUN	PREP	NOUN	.

Unsupervised POS tagging [Schütze, '95]

Part-of-speech (POS) tagging task:

Input:	I	like	reducing	the	dimensionality	of	data	.
Output:	NOUN	VERB	VERB(-ING)	DET	NOUN	PREP	NOUN	.

Key idea: words appearing in similar contexts
should have the same POS tags

Problem: contexts are too sparse

Unsupervised POS tagging [Schütze, '95]

Part-of-speech (POS) tagging task:

Input:	I	like	reducing	the	dimensionality	of	data	.
Output:	NOUN	VERB	VERB(-ING)	DET	NOUN	PREP	NOUN	.

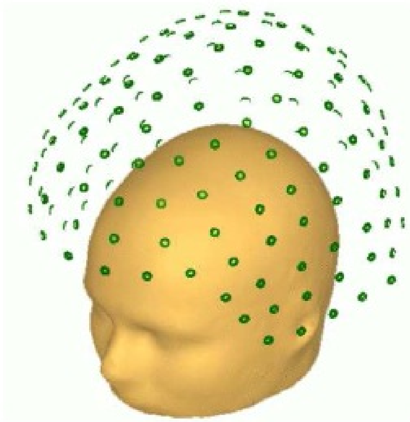
Key idea: words appearing in similar contexts
should have the same POS tags

Problem: contexts are too sparse

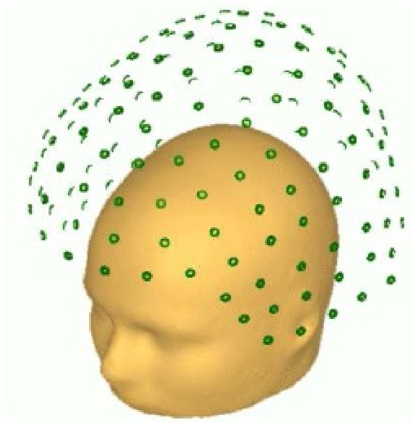
Solution: run PCA first,
then cluster using new representation

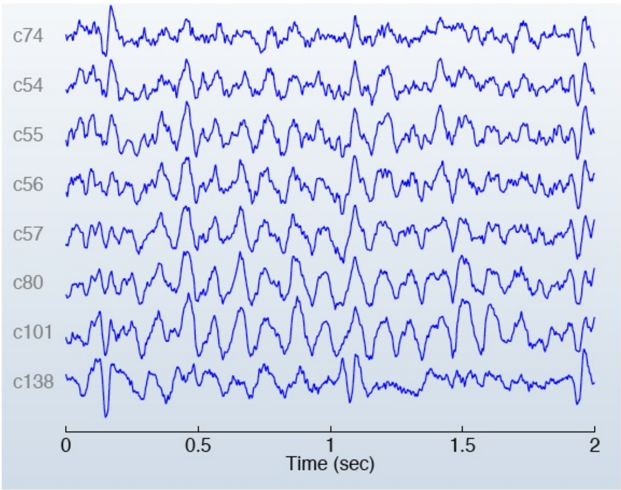
Each data point is (the context of) a word

Brain imaging



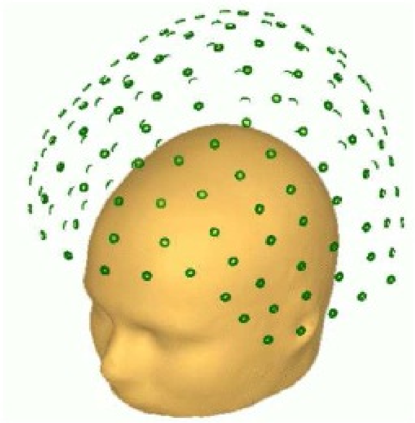
Brain imaging

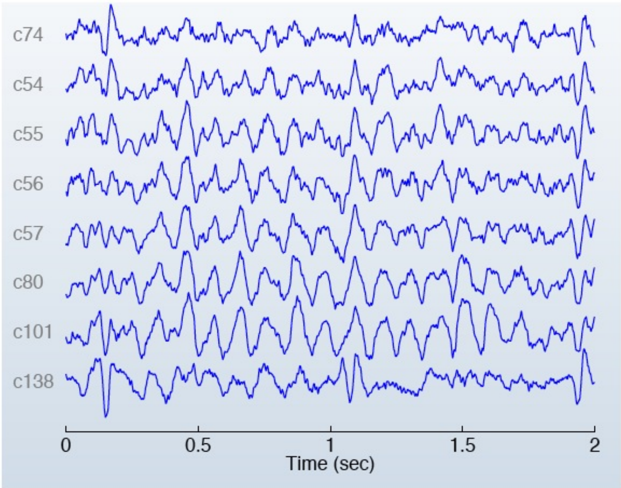


$\mathbf{s} =$ 

Data: EEG/MEG/fMRI readings

Brain imaging



$\mathbf{s} =$ 

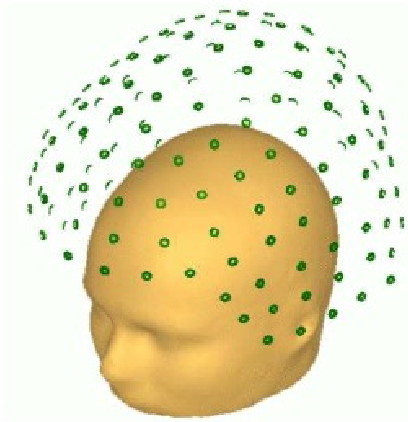
Data: EEG/MEG/fMRI readings

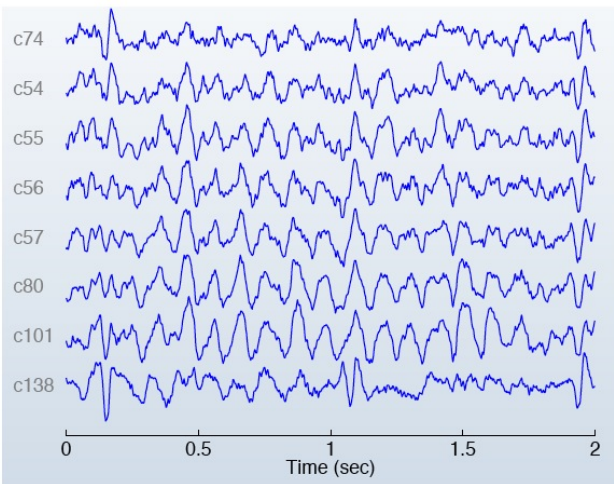
Goal: separate signals into sources

Brain imaging

One solution: ICA

Another solution: CCA [Borga, '02]

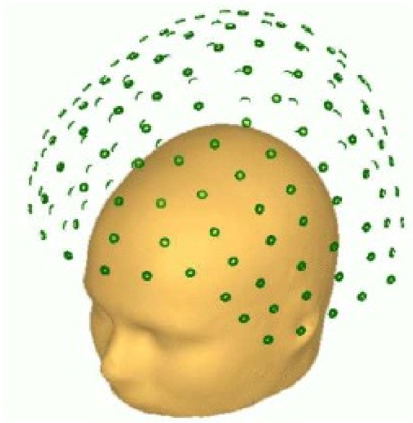


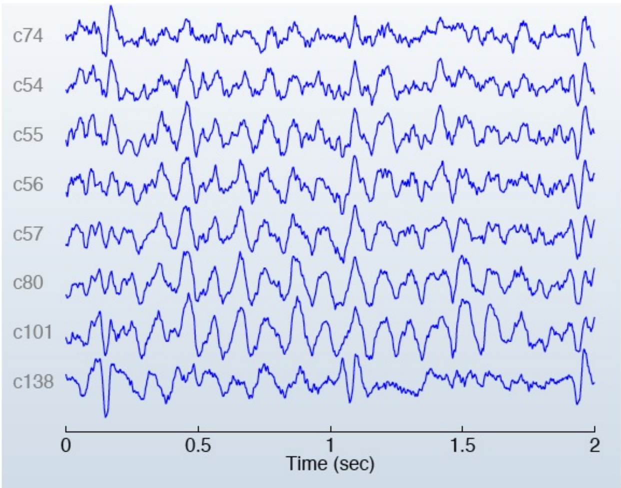
$\mathbf{s} =$ 

Data: EEG/MEG/fMRI readings

Goal: separate signals into sources

Brain imaging



$\mathbf{s} =$ 

Data: EEG/MEG/fMRI readings

Goal: separate signals into sources

One solution: ICA

Another solution: CCA [Borga, '02]

The two views are the signals $\mathbf{s}$ at adjacent time steps:

$$(\mathbf{x}_1, \mathbf{y}_1) = (\mathbf{s}(1), \mathbf{s}(2))$$
$$(\mathbf{x}_2, \mathbf{y}_2) = (\mathbf{s}(2), \mathbf{s}(3))$$
$$(\mathbf{x}_3, \mathbf{y}_3) = (\mathbf{s}(3), \mathbf{s}(4))$$
$$\cdots$$

More robust and faster than ICA

Outline

- Introduction
- Methods
  - Principal component analysis (PCA)
  - Canonical correlation analysis (CCA)
  - Linear discriminant analysis (LDA)
  - Non-negative matrix factorization (NMF)
  - Independent component analysis (ICA)
- Case studies
  - Network anomaly detection
  - Multi-task learning
  - Part-of-speech tagging
  - Brain imaging
- Extensions, related methods, summary

Extensions

- Kernel trick:
  - Find non-linear subspaces with same machinery
- Produce sparse solutions
- Ensure robustness:
  - Be insensitive to outliers
- Make probabilistic (e.g., factor analysis):
  - Handle missing data
  - Estimate uncertainty
  - Natural way to incorporate in a larger model
- Automatically choose number of dimensions

Curtain call

PCA: find subspace that captures most variance in data;
eigenvalue problem

Curtain call

PCA: find subspace that captures most variance in data; eigenvalue problem

CCA: find pair of subspaces that captures most correlation; generalized eigenvalue problem

Curtain call

PCA: find subspace that captures most variance in data;
eigenvalue problem

CCA: find pair of subspaces that captures most correlation;
generalized eigenvalue problem

LDA: find subspace that maximizes $\frac{\text{intraclass variance}}{\text{interclass variance}}$;
generalized eigenvalue problem

Curtain call

PCA: find subspace that captures most variance in data; eigenvalue problem

CCA: find pair of subspaces that captures most correlation; generalized eigenvalue problem

LDA: find subspace that maximizes $\frac{\text{intraclass variance}}{\text{interclass variance}}$; generalized eigenvalue problem

NMF: find subspace that minimizes reconstruction error for non-negative data; non-trivial optimization problem

Curtain call

PCA: find subspace that captures most variance in data; eigenvalue problem

CCA: find pair of subspaces that captures most correlation; generalized eigenvalue problem

LDA: find subspace that maximizes $\frac{\text{intraclass variance}}{\text{interclass variance}}$; generalized eigenvalue problem

NMF: find subspace that minimizes reconstruction error for non-negative data; non-trivial optimization problem

ICA: find subspace where sources are independent; non-trivial optimization problem